

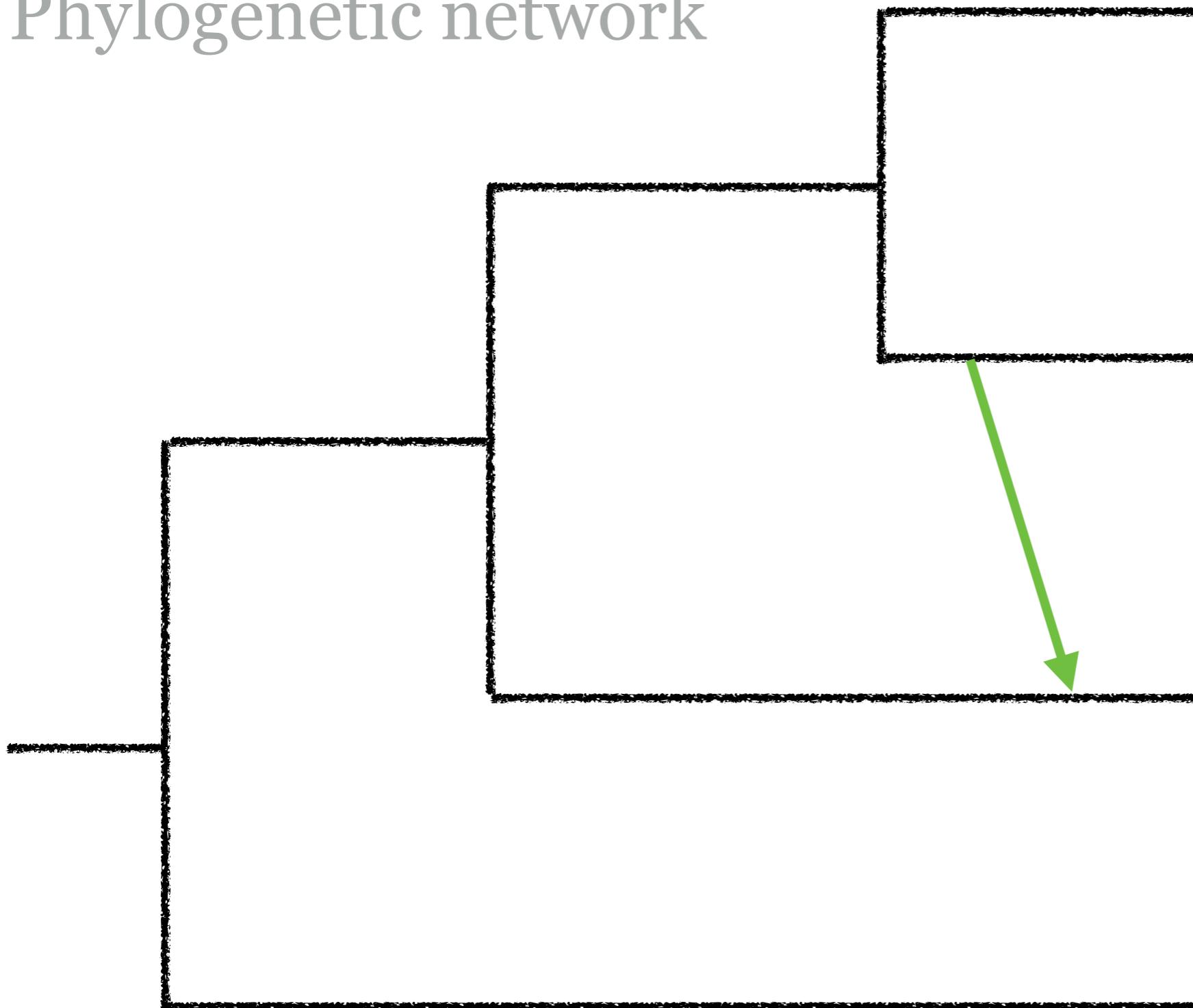
Lecture 14

Coalescent-based methods
Botany 563 – Spring 2022

- **Previous class check-up:**
 - We studied the coalescent model on a species tree
 - We practiced on ASTRAL and/or BUCKY
- **Learning Objectives:** At the end of today's session, you will be able to
 - Explain the coalescent model on a species network
- **Pre-class work**
 - Read SNaQ book chapter

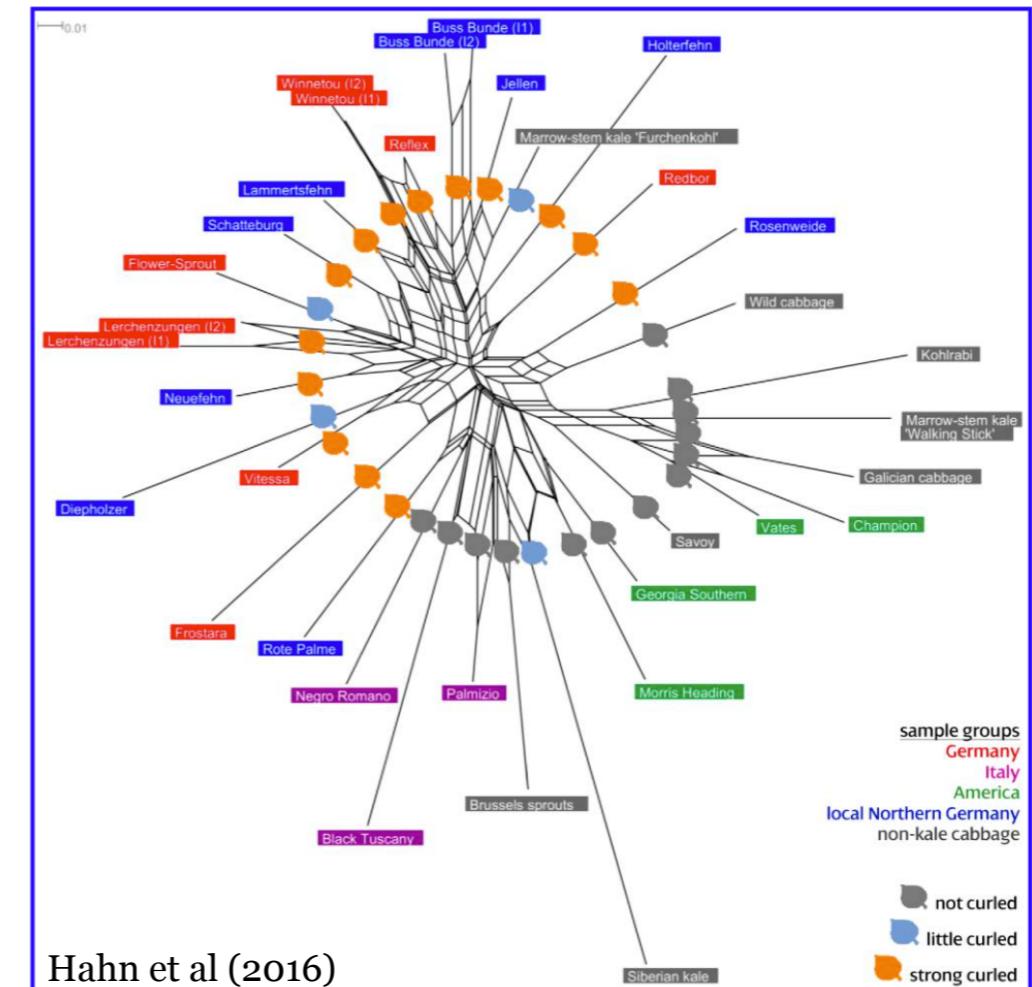
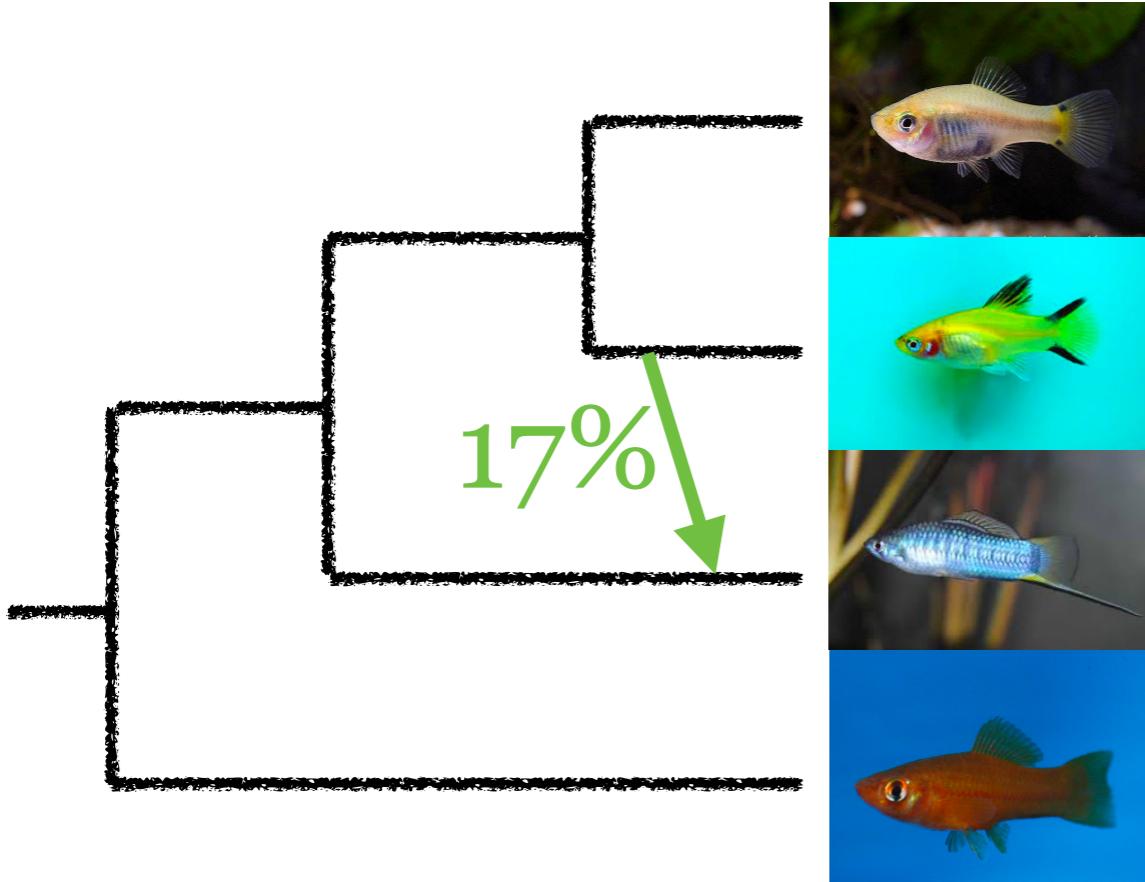
What?

Phylogenetic network



What?

Phylogenetic network

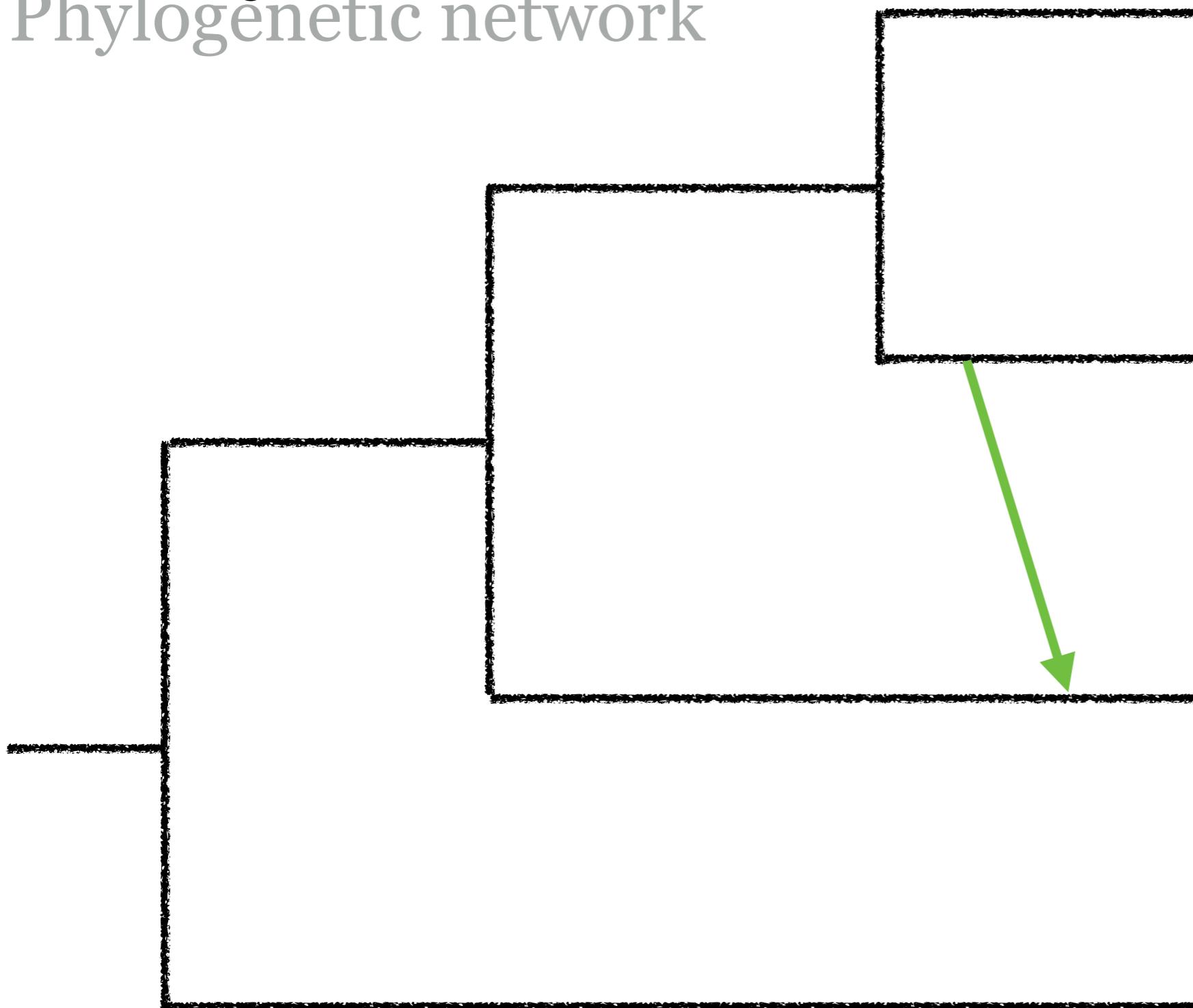


Explicit

Implicit

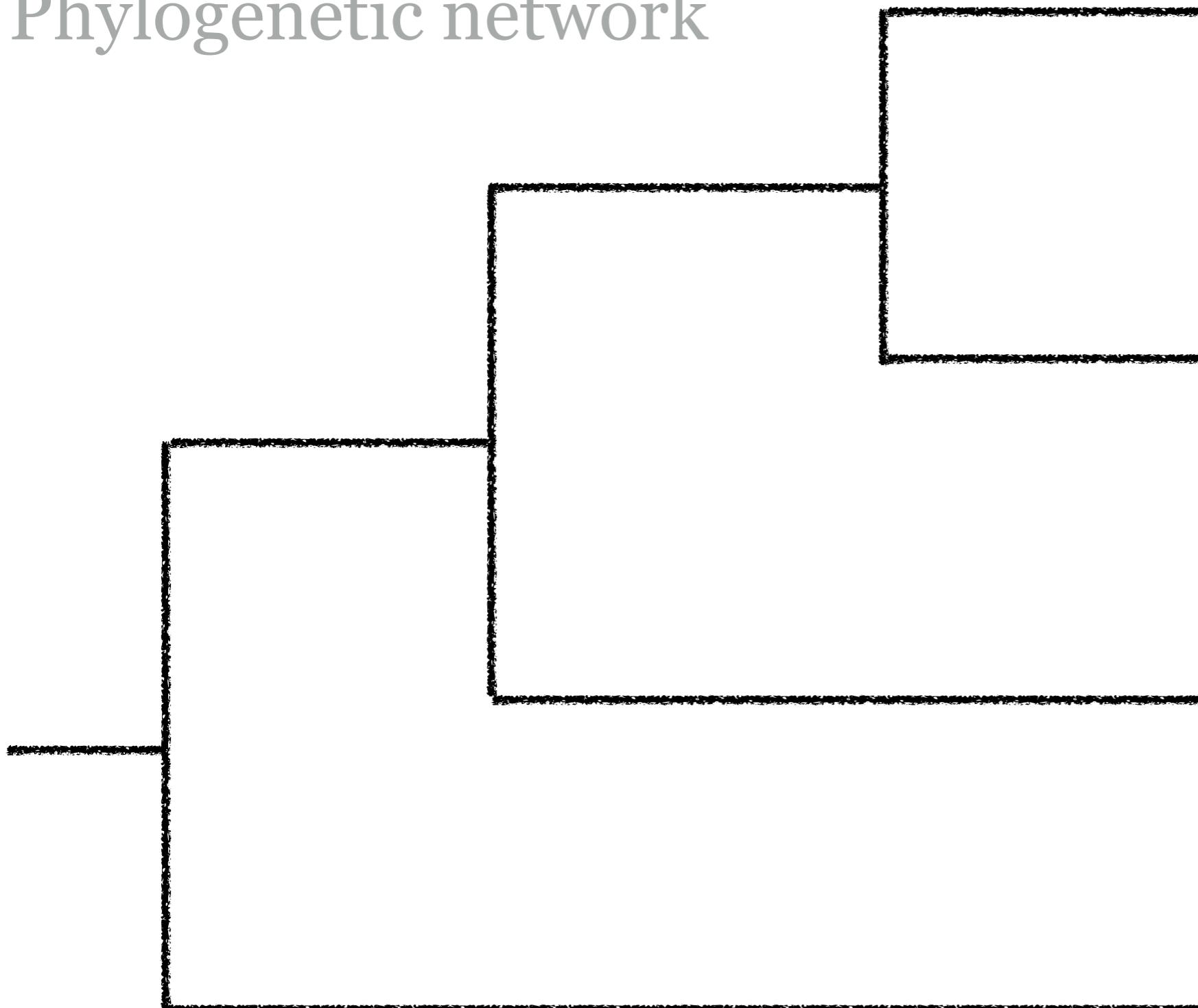
Why?

Phylogenetic network



Why?

Phylogenetic network



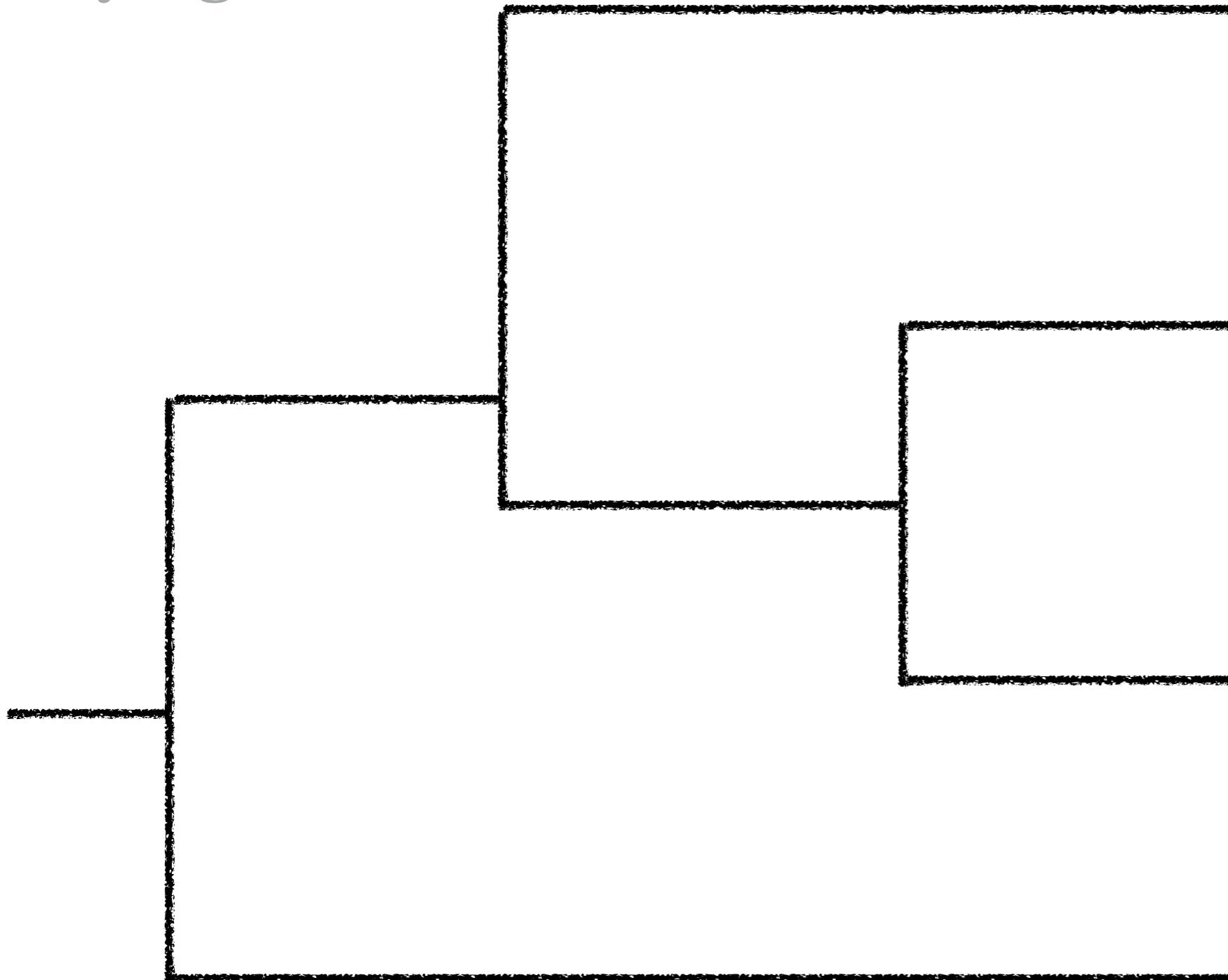
Main tree



Why?

Phylogenetic network

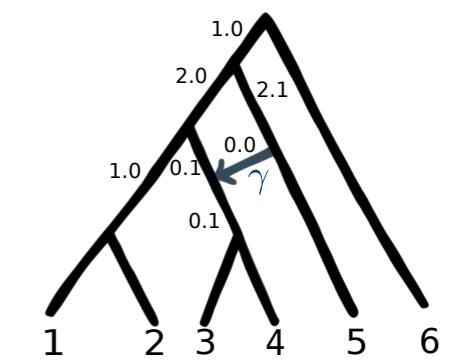
Ignore gene flow
=>Wrong tree!



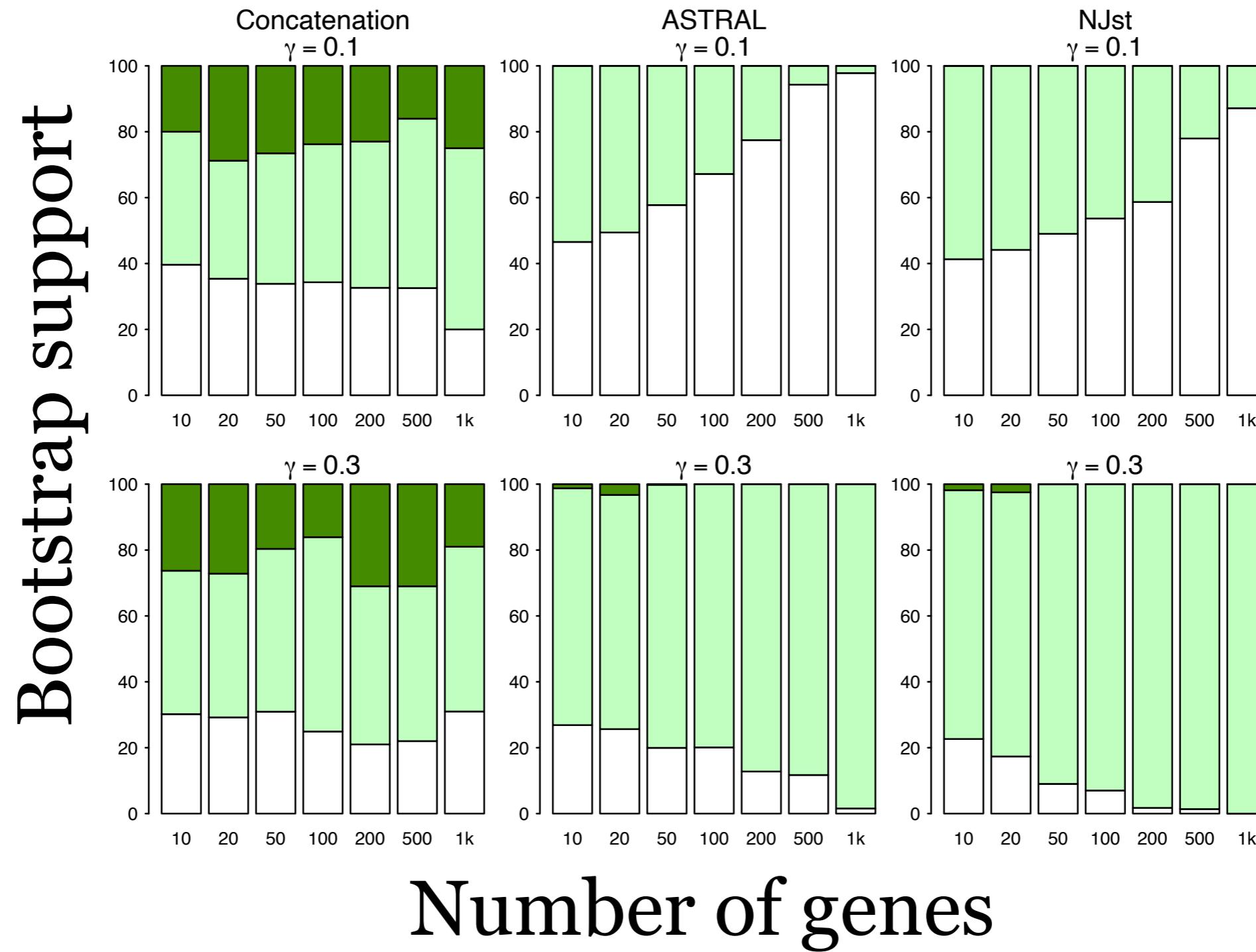
Why?

Phylogenetic network

Coalescent tree methods
not robust to gene flow



White:
true tree

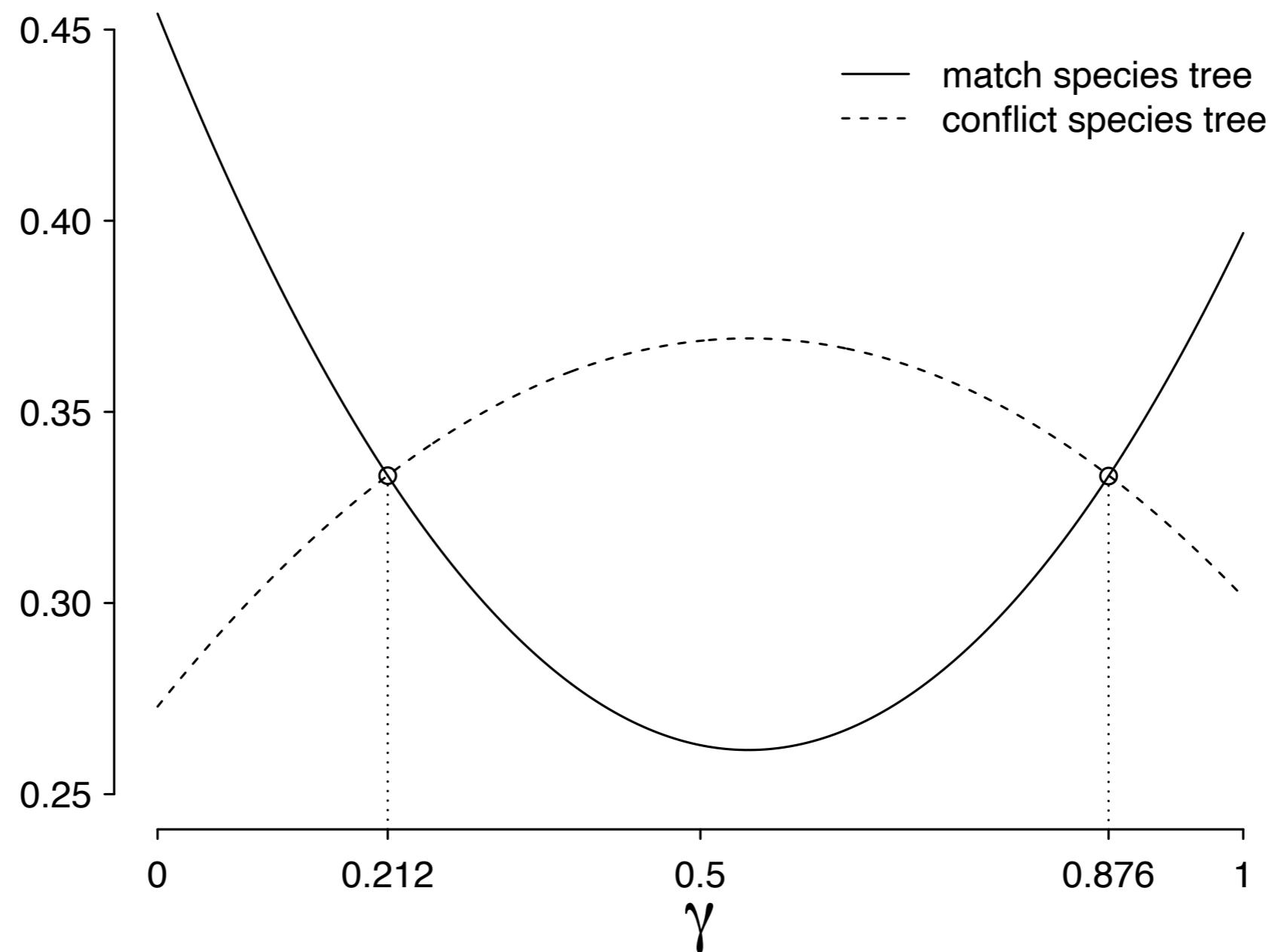


(S.-L., Yang, Ané, 2016, Syst Bio)

ASTRAL (Mirarab et al, 2014)
NJst (Liu&Yu, 2011)

Why? Phylogenetic network

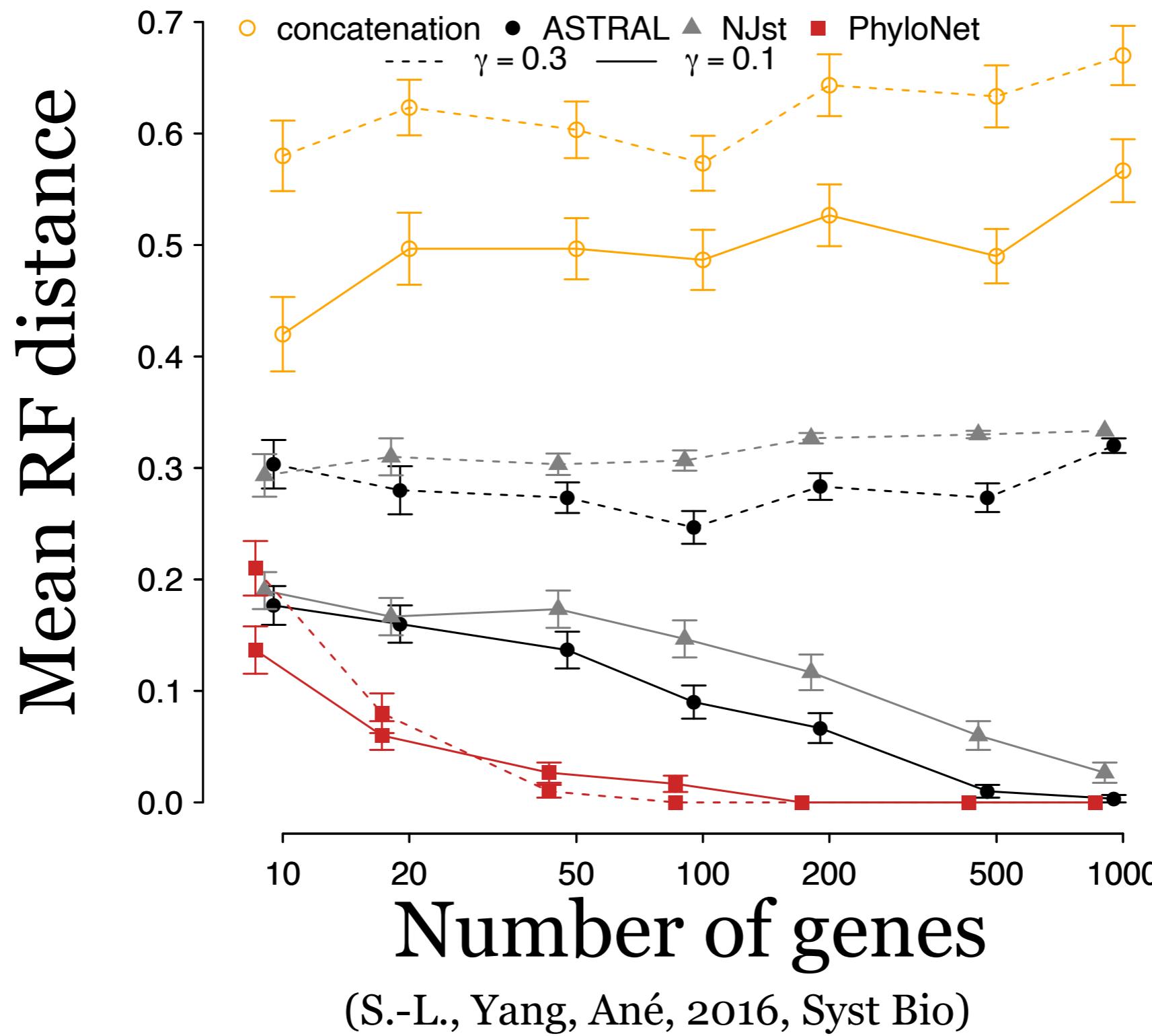
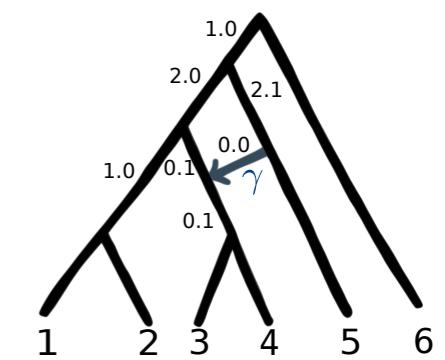
Anomaly zone with gene flow



(S.-L., Yang, Ané, 2016, Syst Bio)

Why? Phylogenetic network

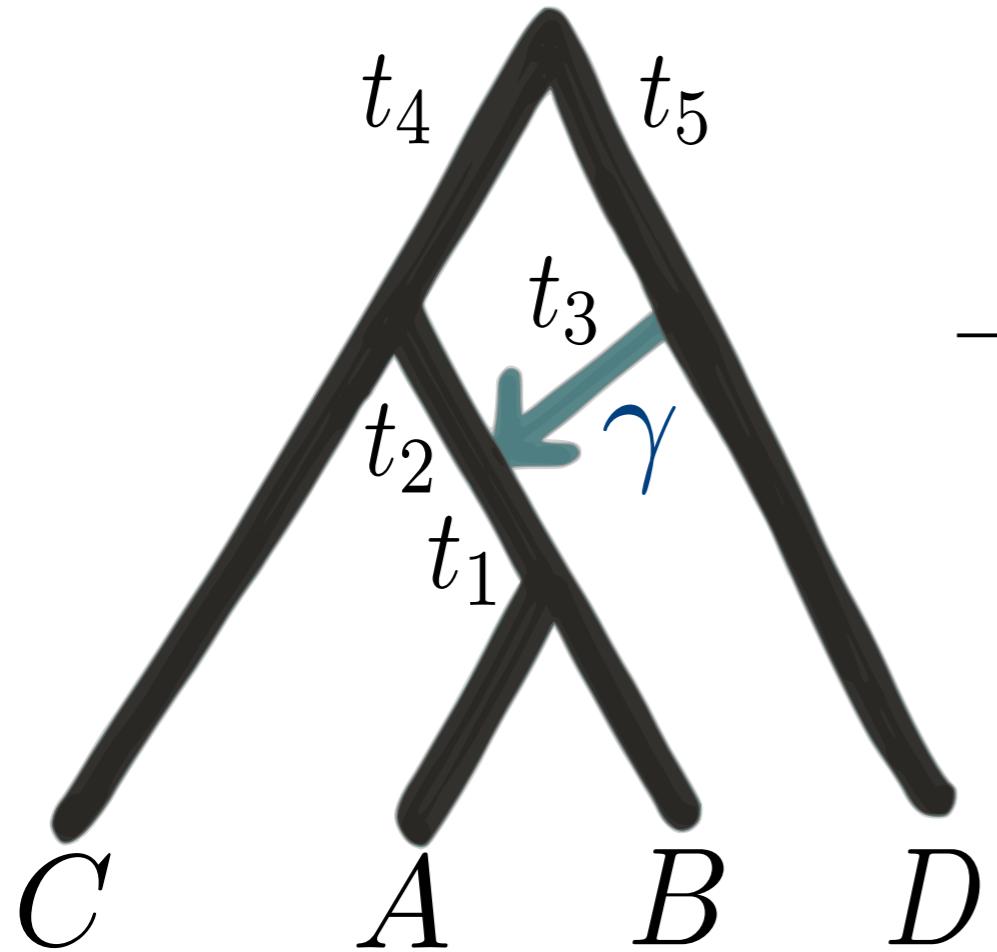
Coalescent tree methods
not robust to gene flow



Why?

Phylogenetic network

Anomalous unrooted
gene trees with gene flow



Frequency among gene trees

Quartet	$\gamma = 0.0$	$\gamma = 0.1$	$\gamma = 0.3$
$AB CD$	0.347	0.298	0.260
$CA BD$	0.327	0.351	0.370
$CB AD$	0.327	0.351	0.370

$$t_1 = t_2 = 0.01, t_3 = t_4 = t_5 = 1$$

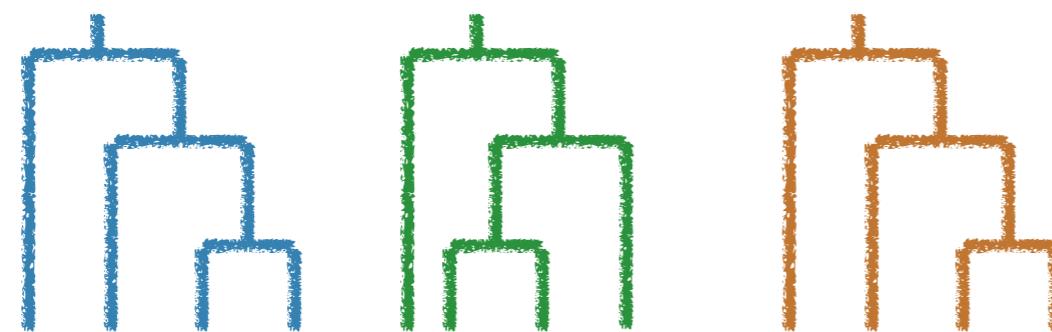
- **ILS**: no AUGT on 4 taxa (Degnan, 2013)
- **ILS+HGT**: AUGT on 4 taxa (S.-L., Yang, Ané, 2016, Syst Bio)

How?

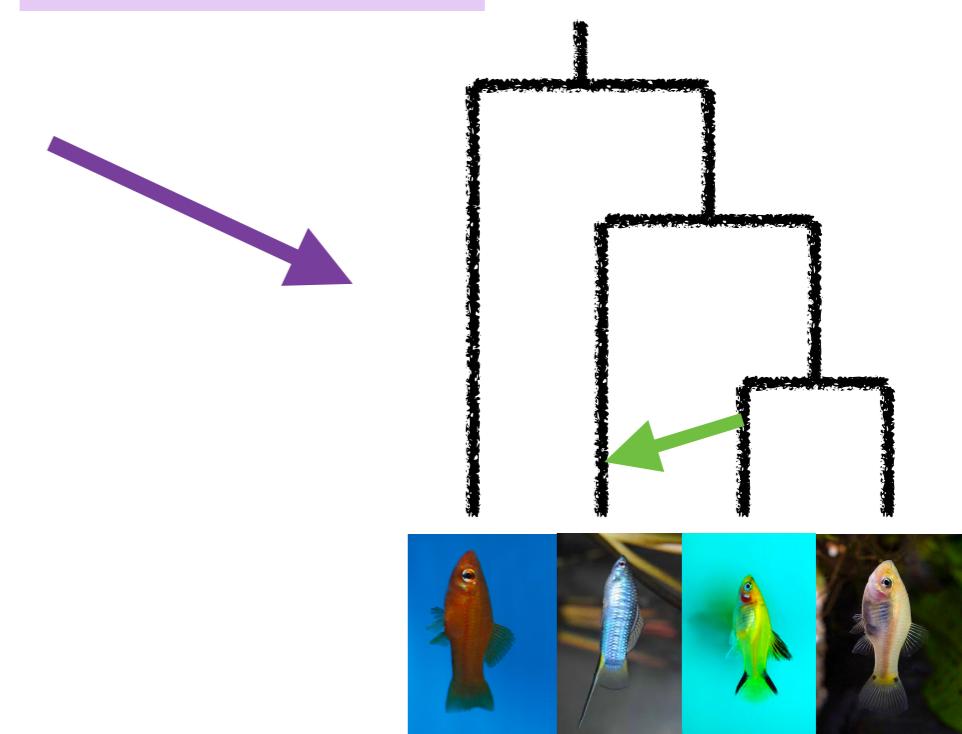
Phylogenetic network



MrBayes
(Huelsenbeck, Ronquist, 2001)
RAxML
(Stamatakis, 2014)
PhyML
(Guindon et al, 2010)



BEAST2
(Zhang et al, 2017)
PhyloNet
(Wen et al, 2016)



SNaQ
(S.-L., Ane, 2016)
PhyloNet
(Yu et al, 2014)

Multispecies coalescent on a network



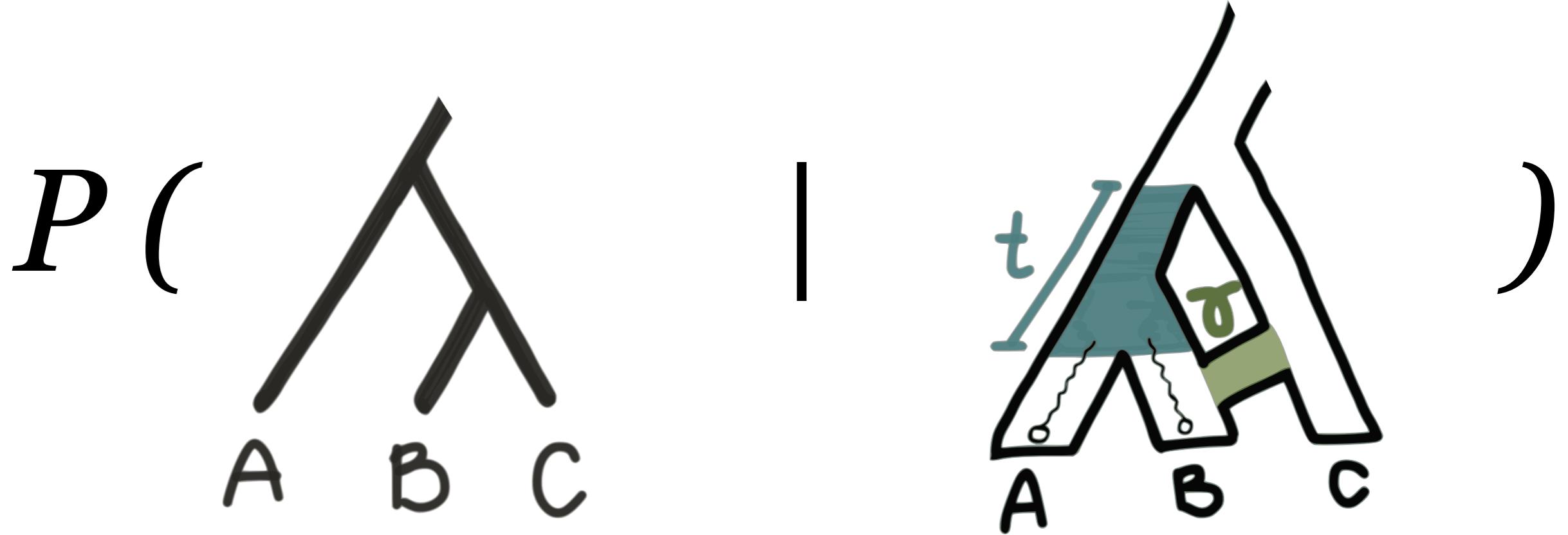
(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)

Multispecies coalescent on a network



(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)

Multispecies coalescent on a network



(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)



<https://solislemuslab.github.io/>

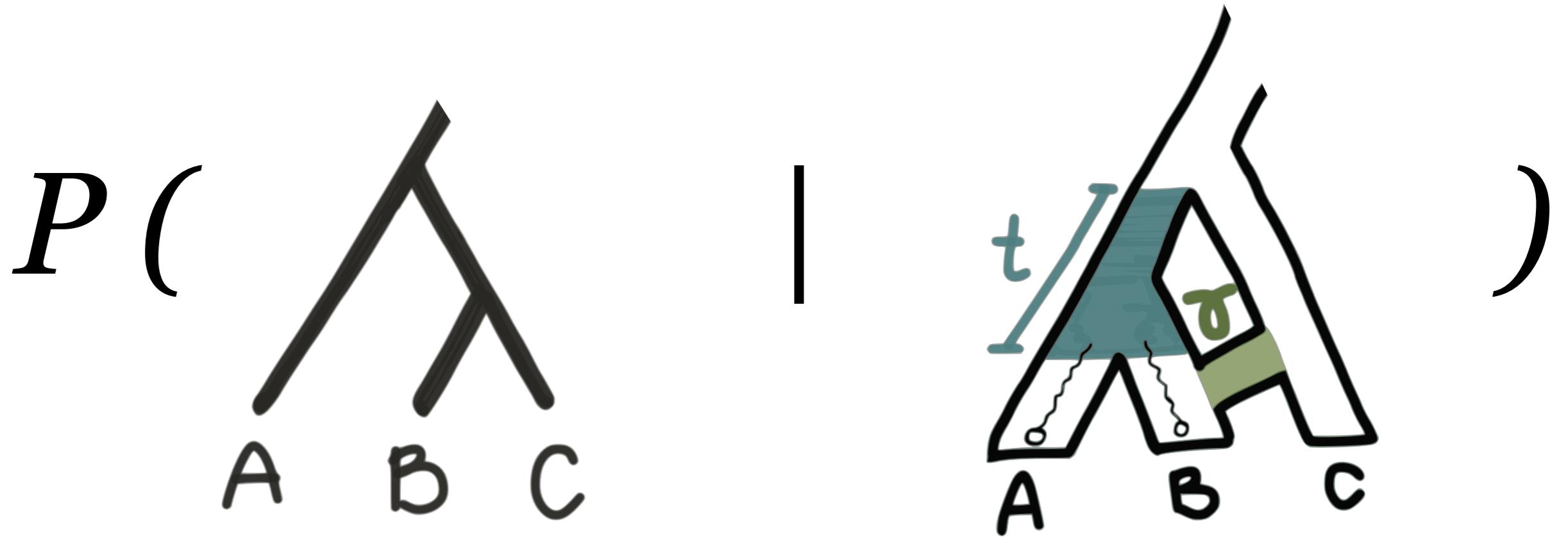


@solislemuslab



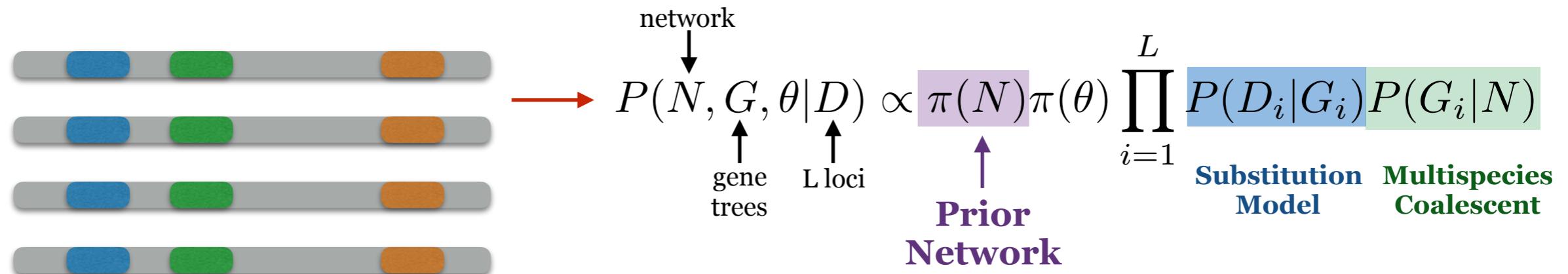
crsl4

Multispecies coalescent on a network



$$p_{BC|AD}(t, t_2, \gamma) = (1 - \gamma) \frac{1}{3} e^{-t} + \gamma (1 - \frac{2}{3} e^{-t_2})$$

(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)

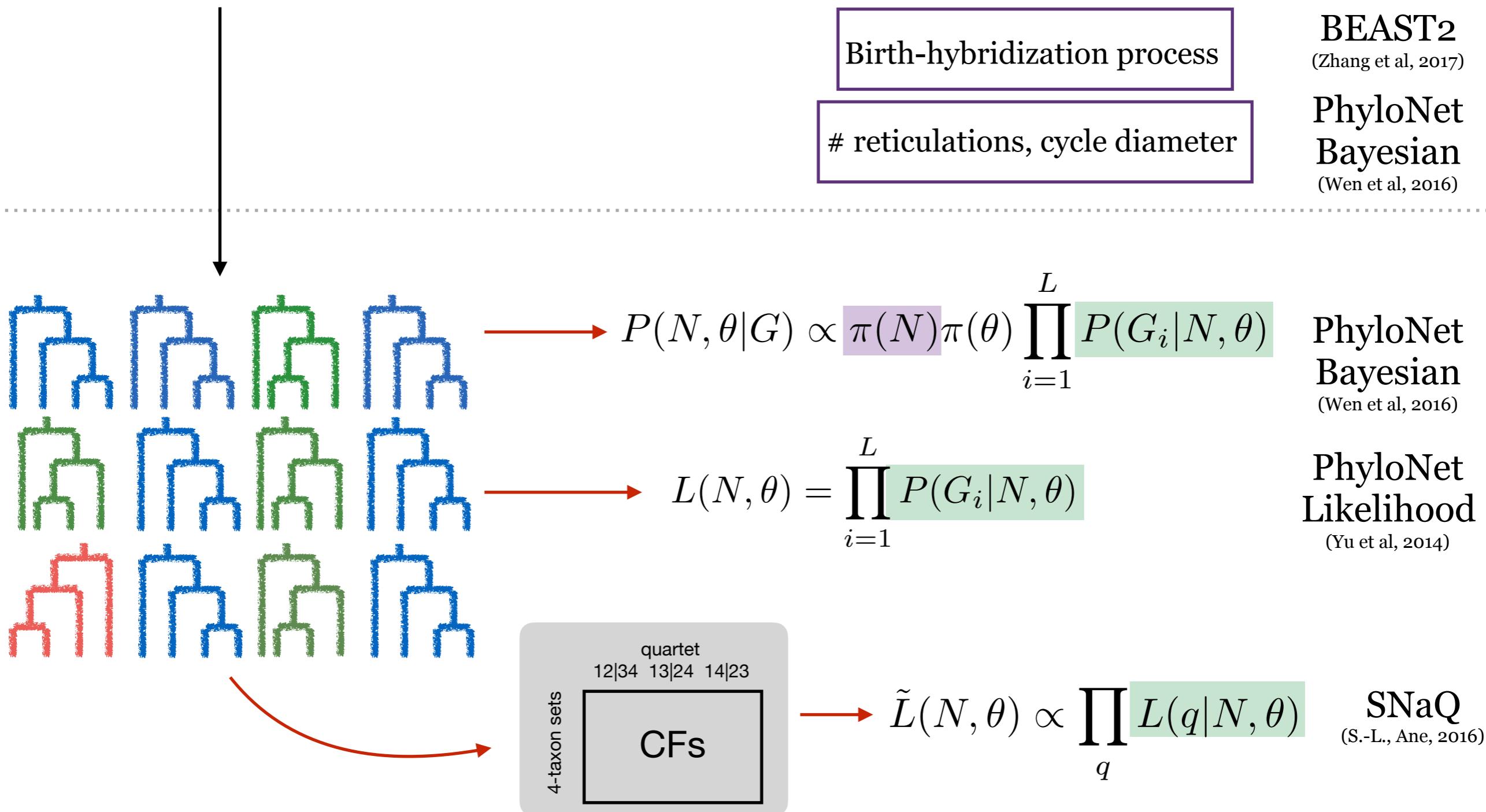
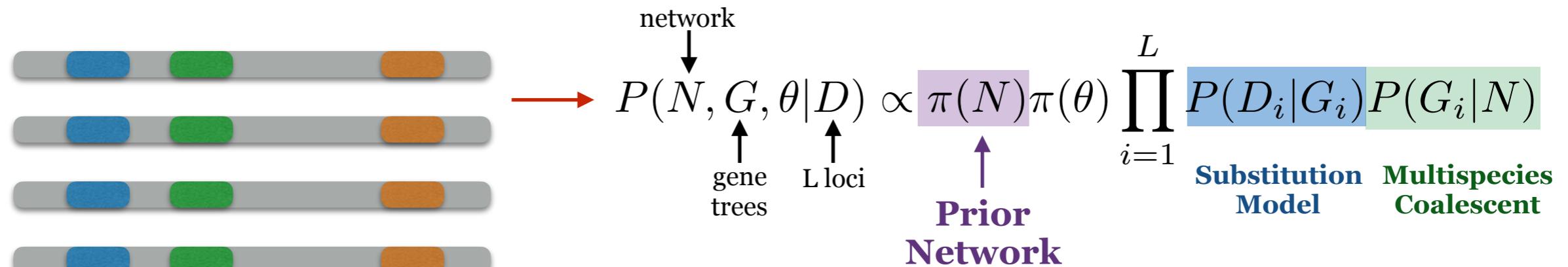


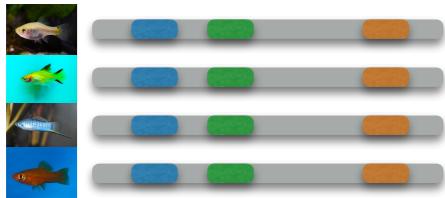
Birth-hybridization process

reticulations, cycle diameter

BEAST2
(Zhang et al, 2017)

PhyloNet Bayesian
(Wen et al, 2016)

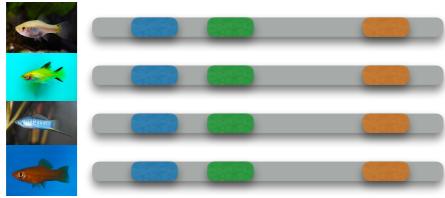




BEAST2
(Zhang et al, 2017)

Birth-hybridization process

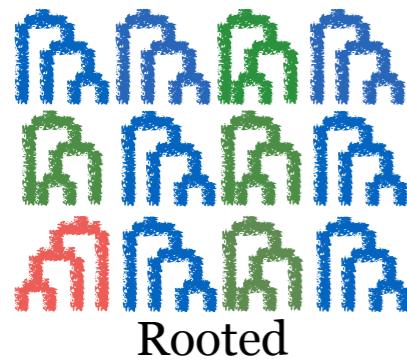
Most accurate,
not scalable



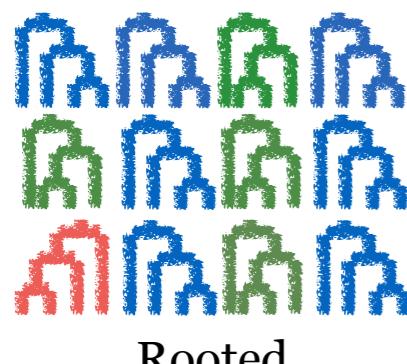
PhyloNet
Bayesian
(Wen et al, 2016)

MCMC:
Network moves,
mixing

reticulations,
cycle diameter

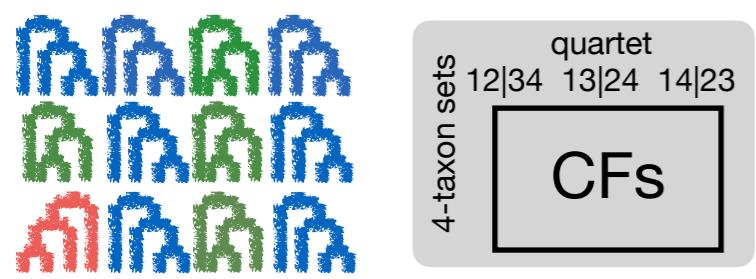


PhyloNet
Bayesian
(Wen et al, 2016)



PhyloNet
Likelihood
(Yu et al, 2014)

Heuristic search:
Network moves



SNaQ
(S.-L., Ane, 2016)

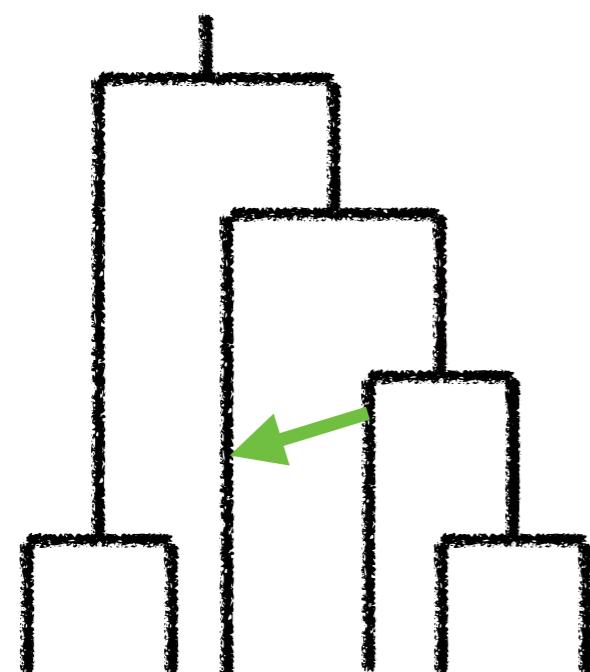
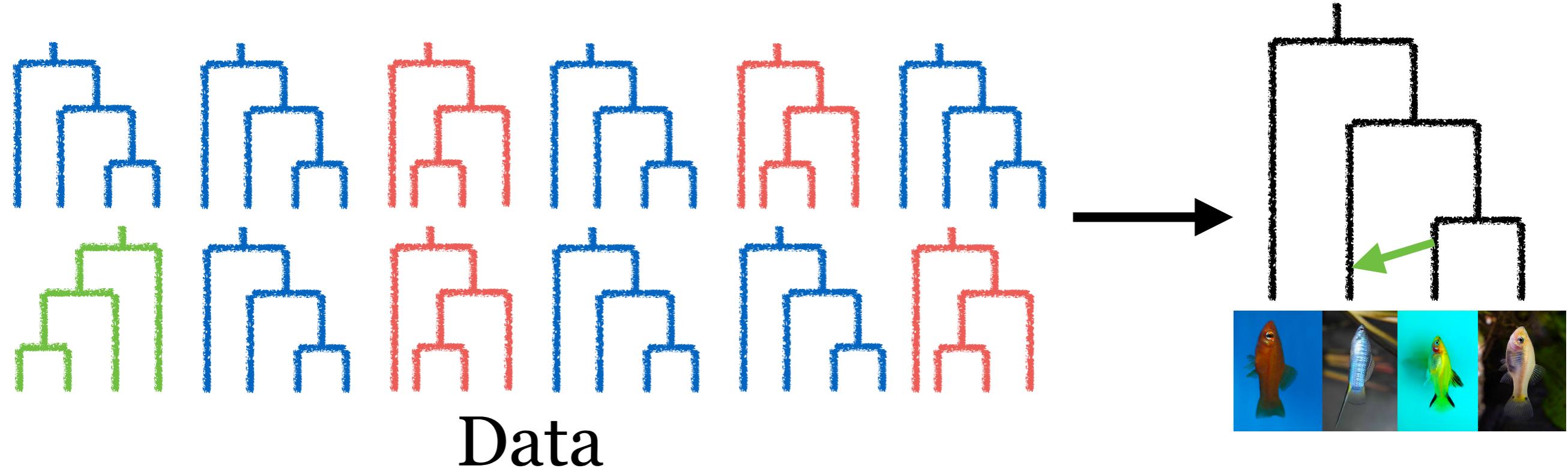
Level-1
networks

More scalable,
Robust

Unrooted

STEM-hy	gene trees rooted, BL	likelihood	hybridization b/w sister lineages
PhyloNet InferNetwork_ML	gene trees rooted	likelihood	
PhyloNet InferNetwork_MPL	gene trees rooted	triplet likelihood	
Phylogenetworks SNaQ	gene trees or quartet CFs	quartet likelihood	level-1 network
PhyloNet MCMC_GT	gene trees rooted	Bayesian	compound prior
PhyloNet MCMC_SEQ	alignments	Bayesian	compound prior no rate variation
BEAST2 SpeciesNetwork	alignments	Bayesian	birth-hyb prior
PhyloNet MLE_BiMarkers	biallelic sites	likelihood	compound prior
PhyloNet MCMC_BiMarkers	biallelic sites	Bayesian	compound prior
HyDe	sites	invariants	4 taxa, 1 hyb.

Maximum pseudolikelihood



Quartet-based inference

$$\tilde{L}(\textit{network}) = \prod L(\textit{quartet})$$

(S-L, Ané, 2016, PLoS Genetics)

www.github.com/CRSL4/PhyloNetworks

snaQ julia



<https://solislemuslab.github.io/>

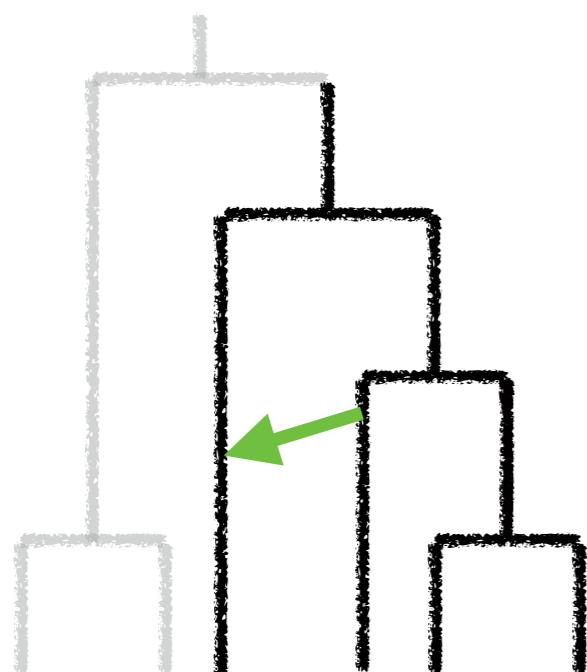
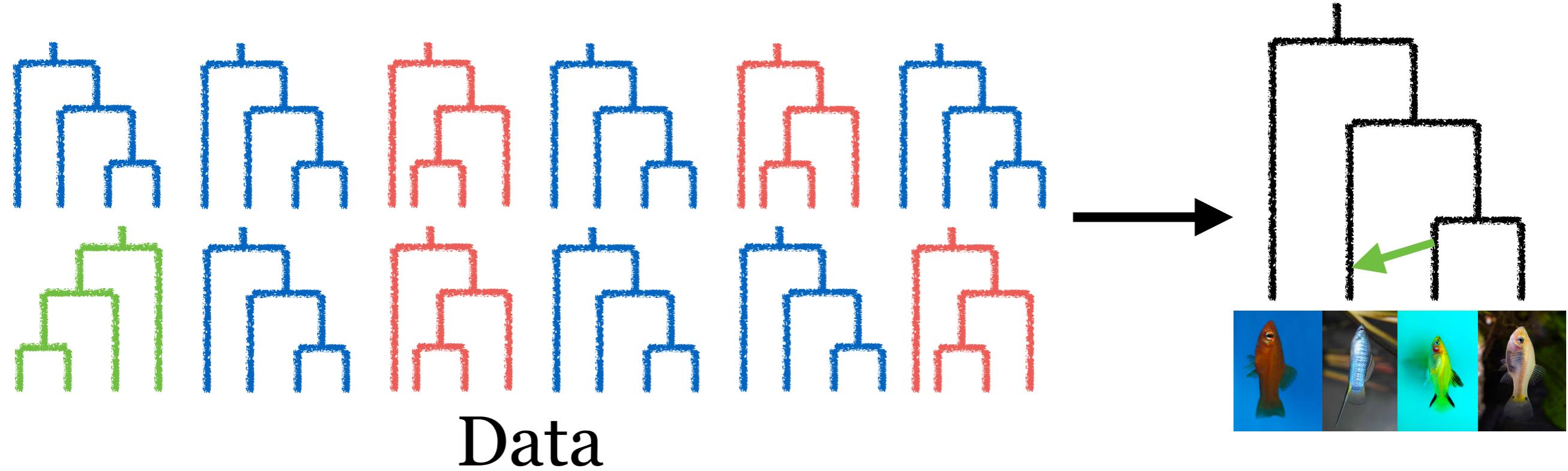


@solislemuslab



crsl4

Maximum pseudolikelihood



Quartet-based inference

$$\tilde{L}(\textit{network}) = \prod L(\textit{quartet})$$

(S-L, Ané, 2016, PLoS Genetics)

[www.github.com/CRSL4/PhyloNetworks](https://github.com/CRSL4/PhyloNetworks)

snaQ julia



<https://solislemuslab.github.io/>

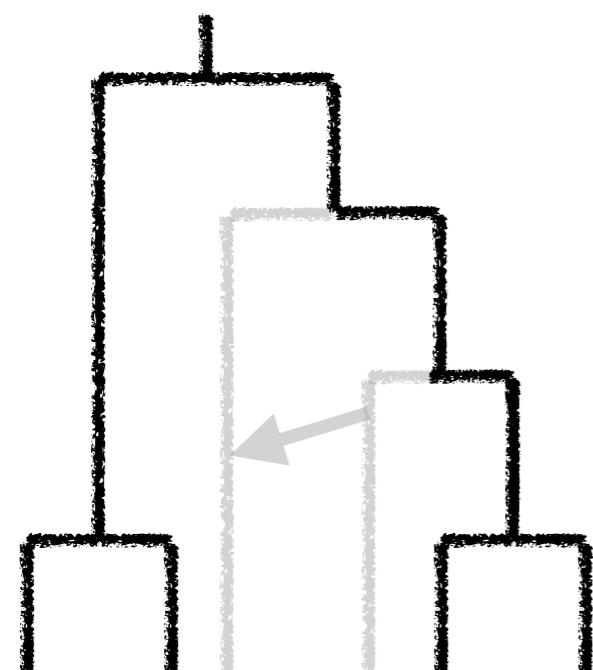
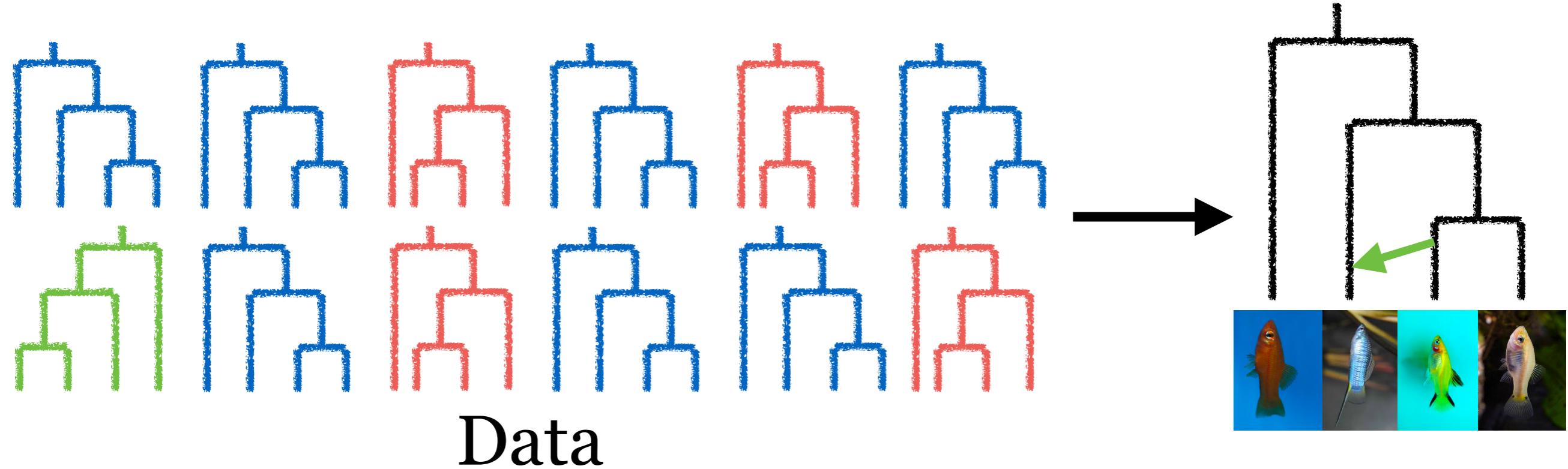


@solislemuslab



crsl4

Maximum pseudolikelihood



Quartet-based inference

$$\tilde{L}(\textit{network}) = \prod L(\textit{quartet})$$

(S-L, Ané, 2016, PLoS Genetics)

[www.github.com/CRSL4/PhyloNetworks](https://github.com/CRSL4/PhyloNetworks)

snaQ julia



<https://solislemuslab.github.io/>



@solislemuslab



crsl4

Maximum pseudolikelihood

Unrooted gene trees

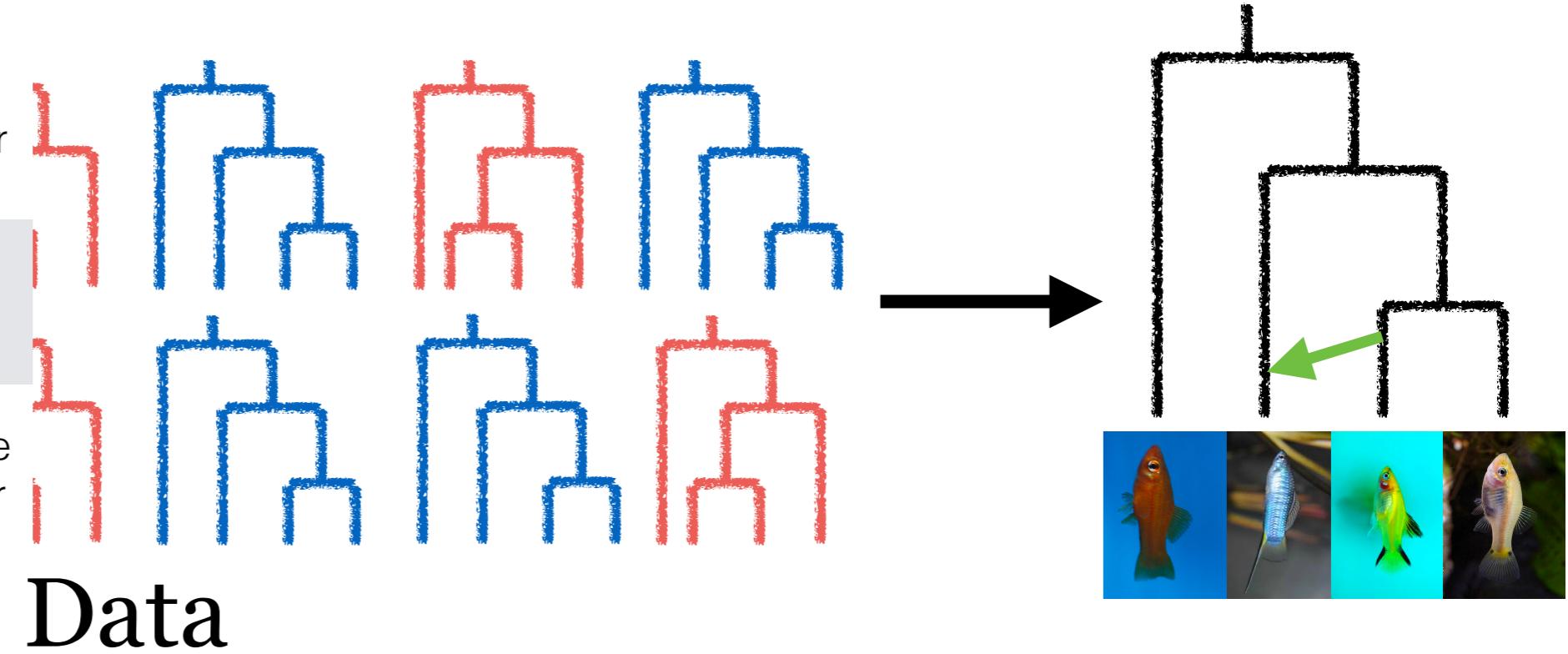
No branch lengths

Concordance factors

No rooting error

No molecular clock assumption

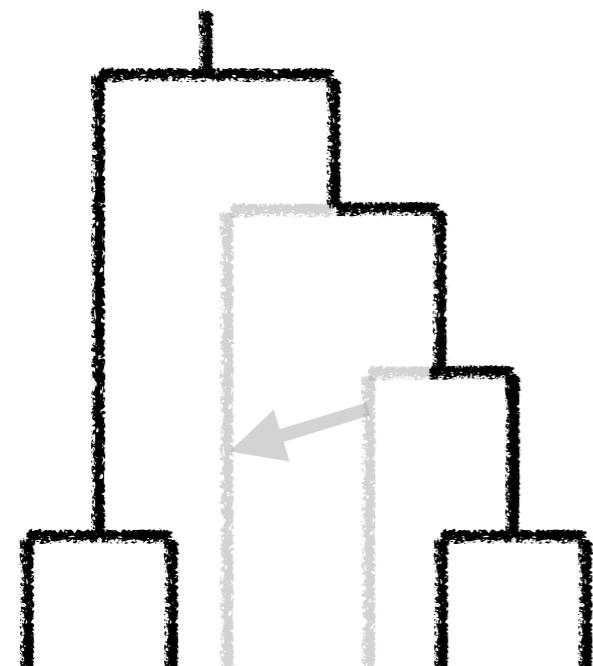
Account for tree estimation error



$$\tilde{L}(\textit{network}) = \prod L(\textit{quartet})$$

(S-L, Ané, 2016, PLoS Genetics)

[www.github.com/CRSL4/PhyloNetworks](https://github.com/CRSL4/PhyloNetworks)



Quartet-based inference

snaQ julia



<https://solislemuslab.github.io/>

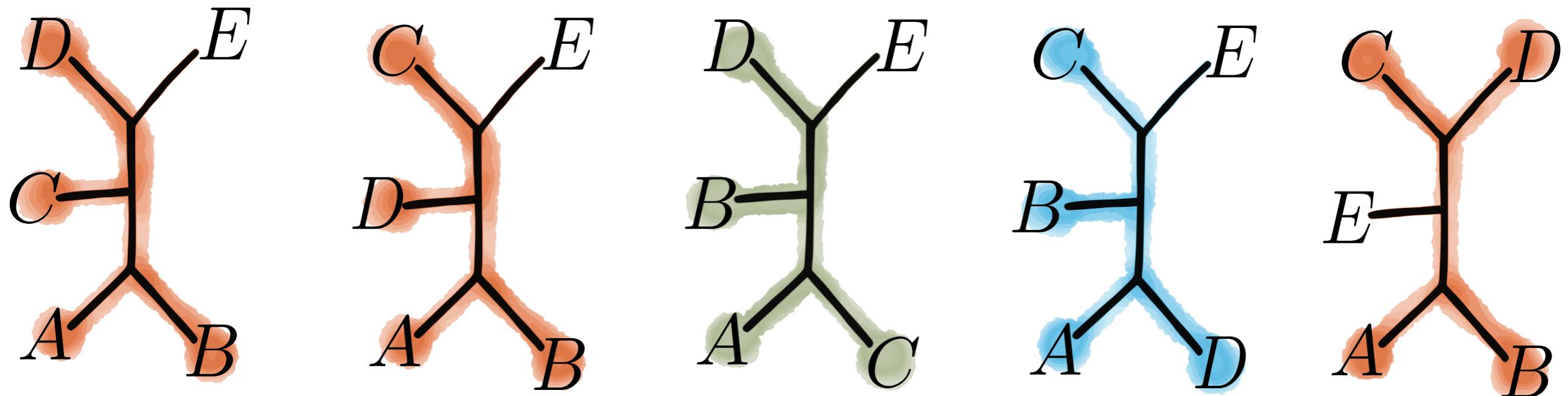


@solislemuslab

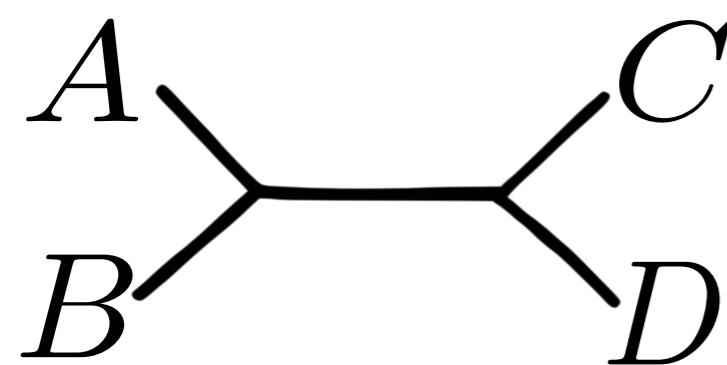


crsl4

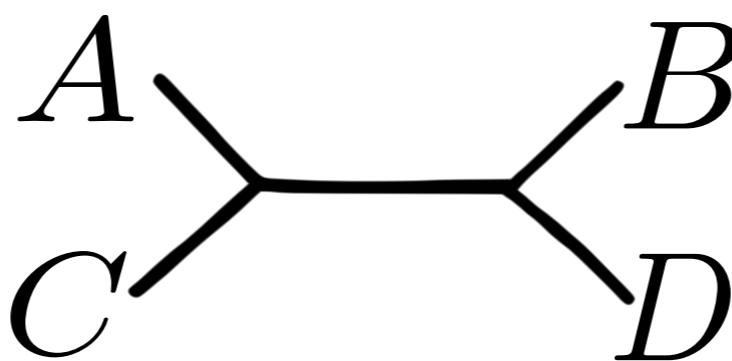
Quartet-based inference



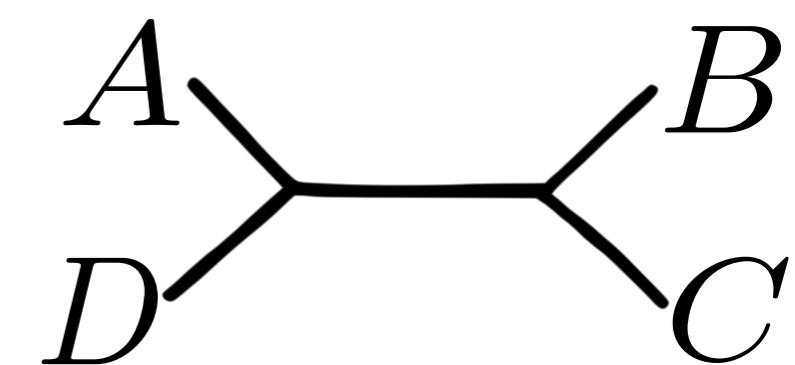
Concordance factors (CF):
% of genes having the quartet in their tree



3/5



1/5



1/5



<https://solislemuslab.github.io/>



@solislemuslab



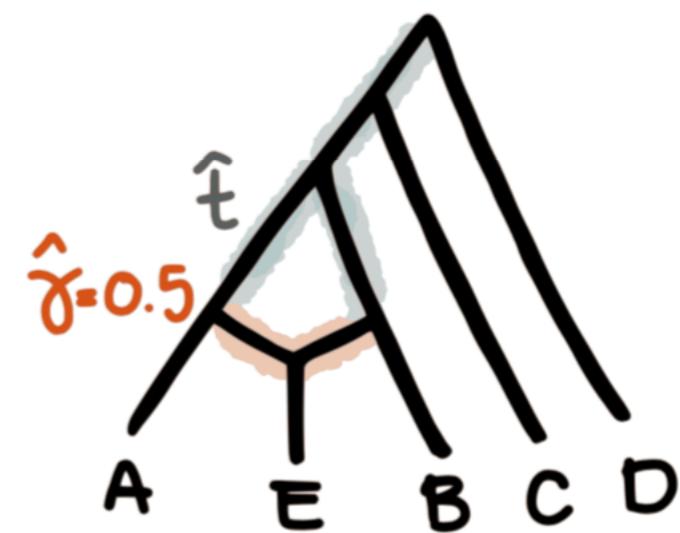
crsl4

Quartet-based inference

Observed **quartet** CFs:

4 taxon set	CF_1	CF_2	CF_3
A B C D	.80	.10	.10
A B C E	.40	.40	.20
A B D E	.40	.40	.20
A C D E	.84	.08	.08
B C D E	.82	.10	.08

inferred network:

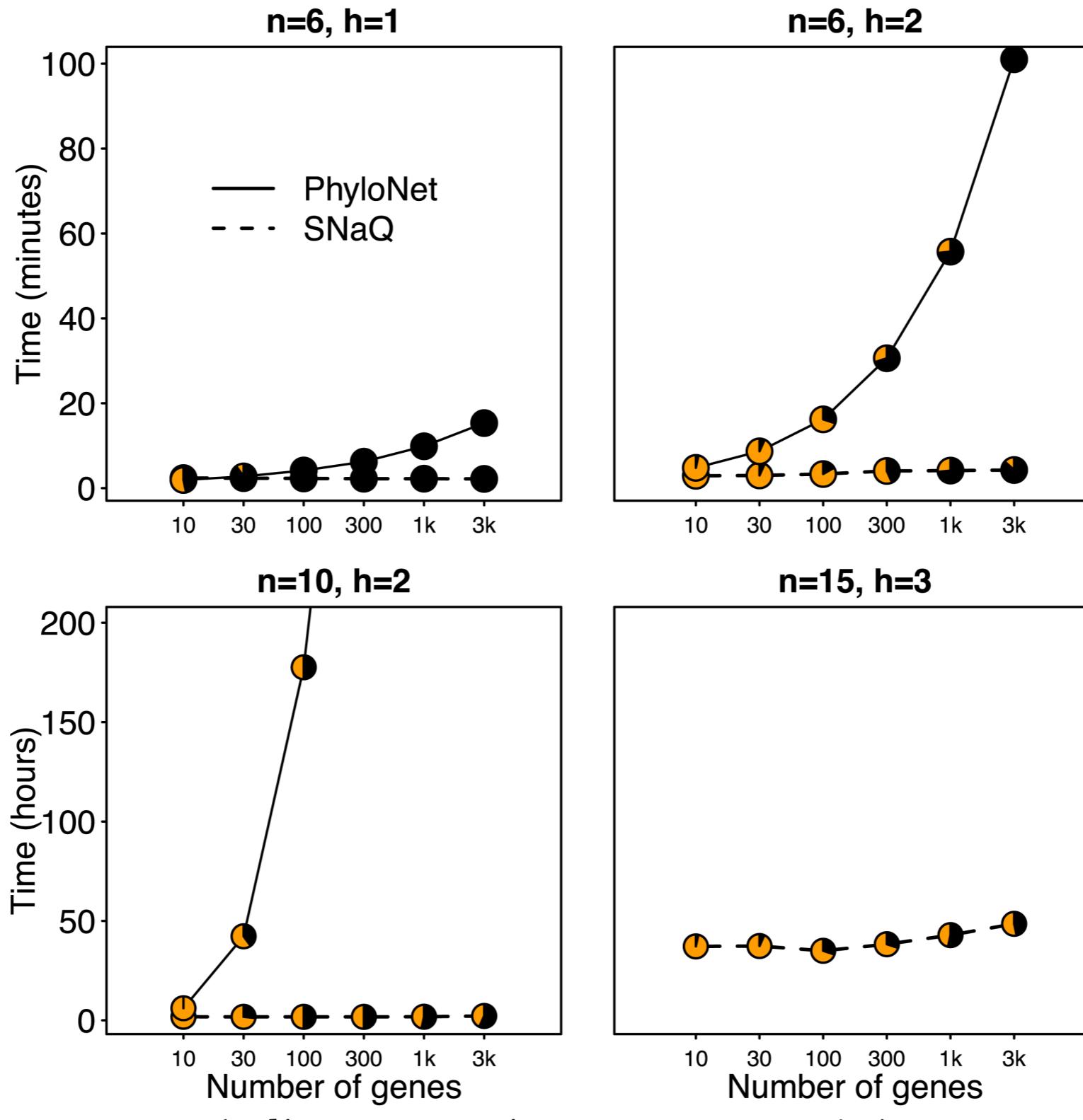


Maximum Pseudo-Likelihood:

$$\log \tilde{L} = \sum_{q \in Q(N)} CF_{in,1} \log(CF_{net,1}) + CF_{in,2} \log(CF_{net,2}) + CF_{in,3} \log(CF_{net,3})$$



Scalability gains



<https://solislemuslab.github.io/>

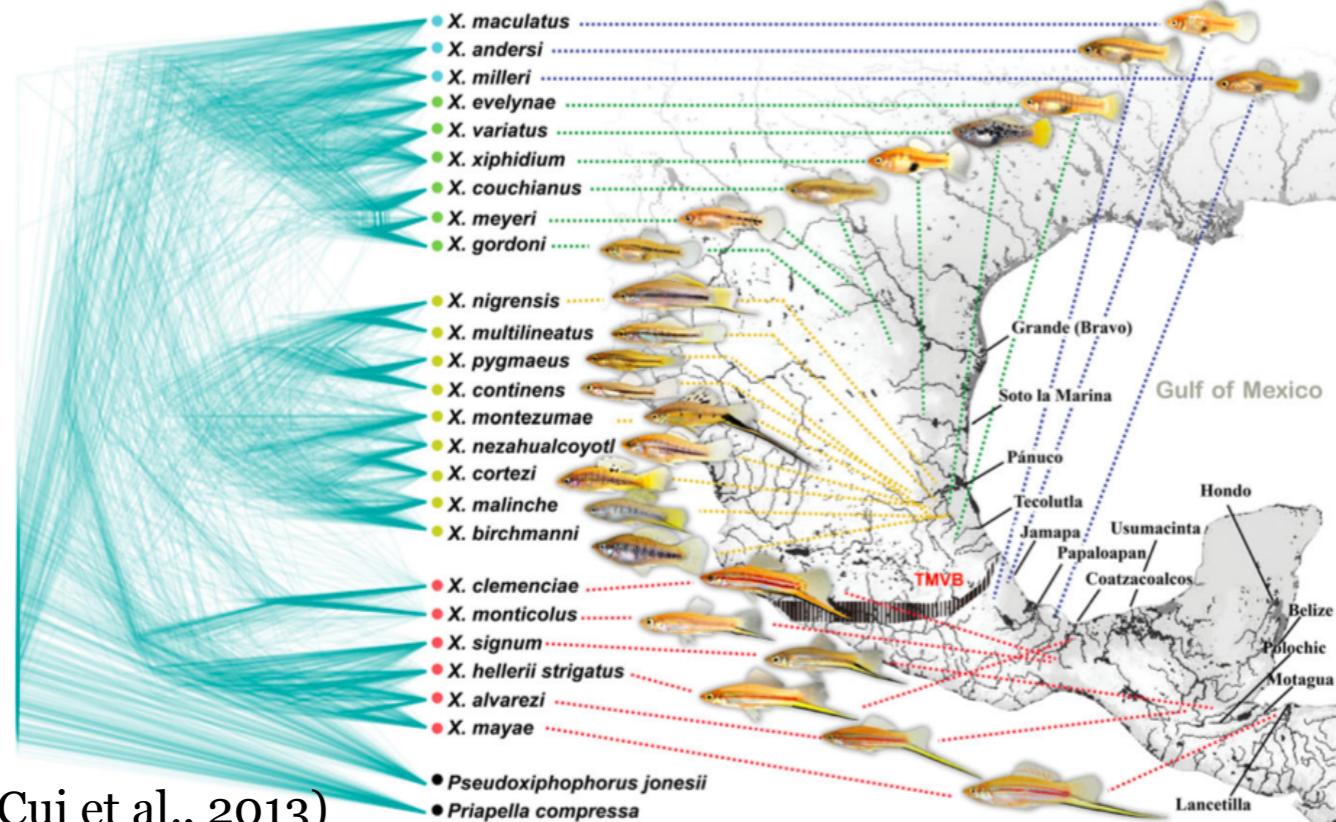


@solislemuslab

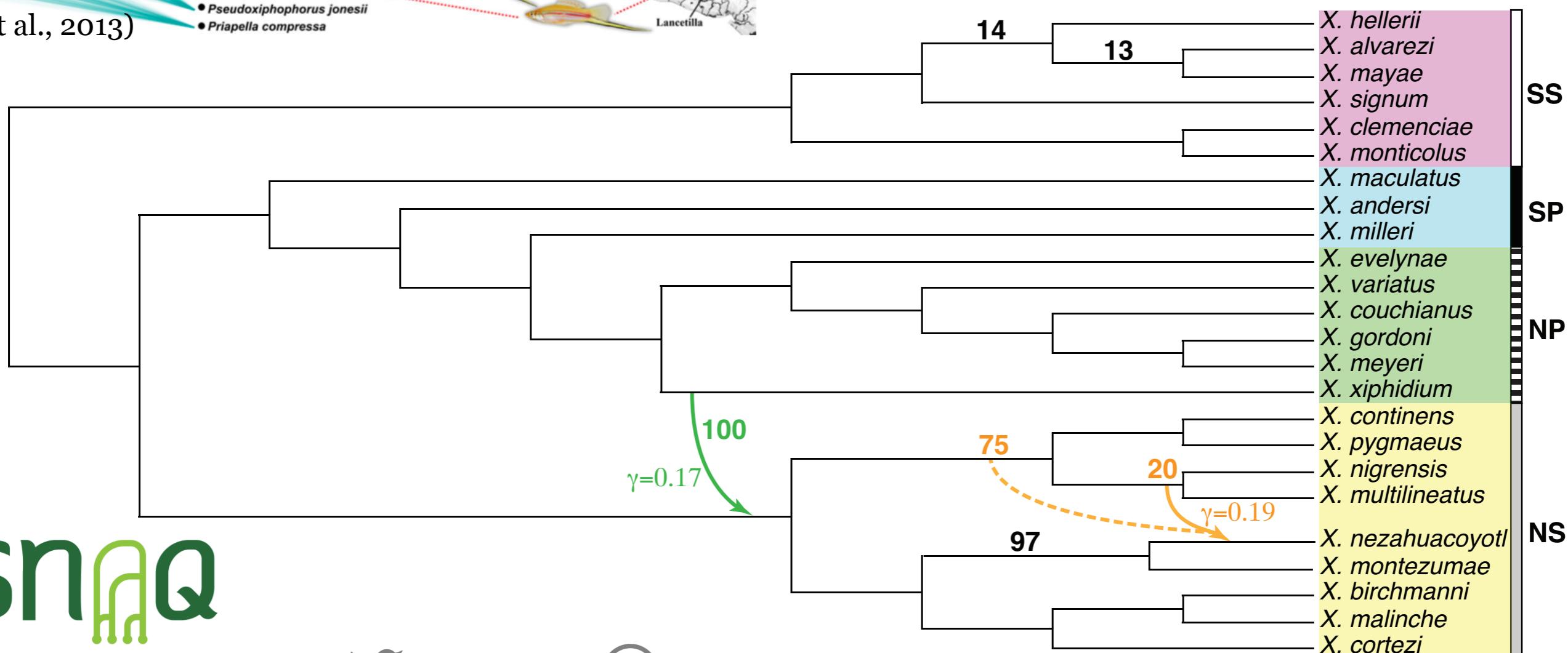


crsl4

1183 genes, 24 swordtails and platyfish



Xiphophorus fish data



snaQ



<http://solislemuslab.github.io/>



@solislemuslab



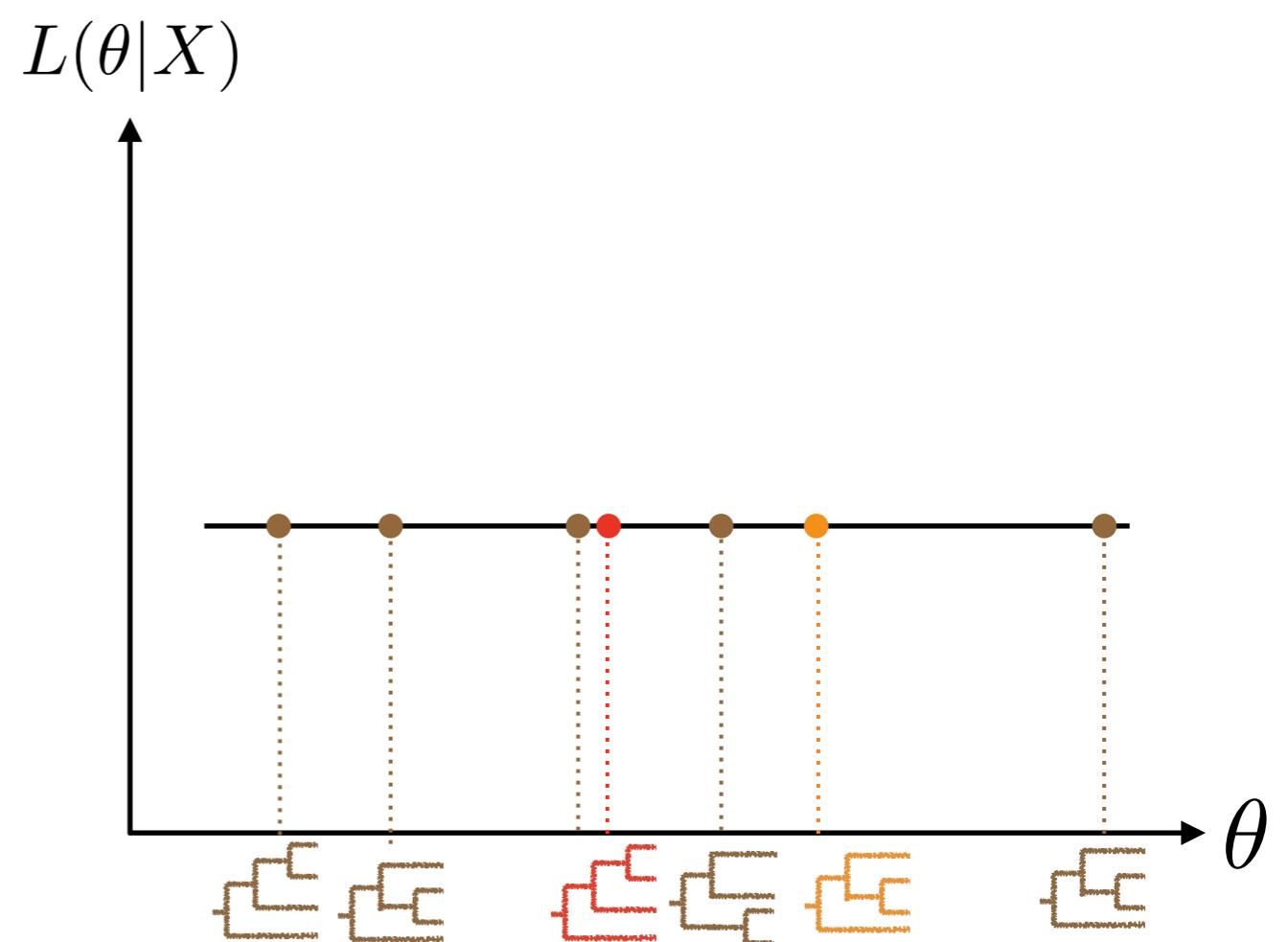
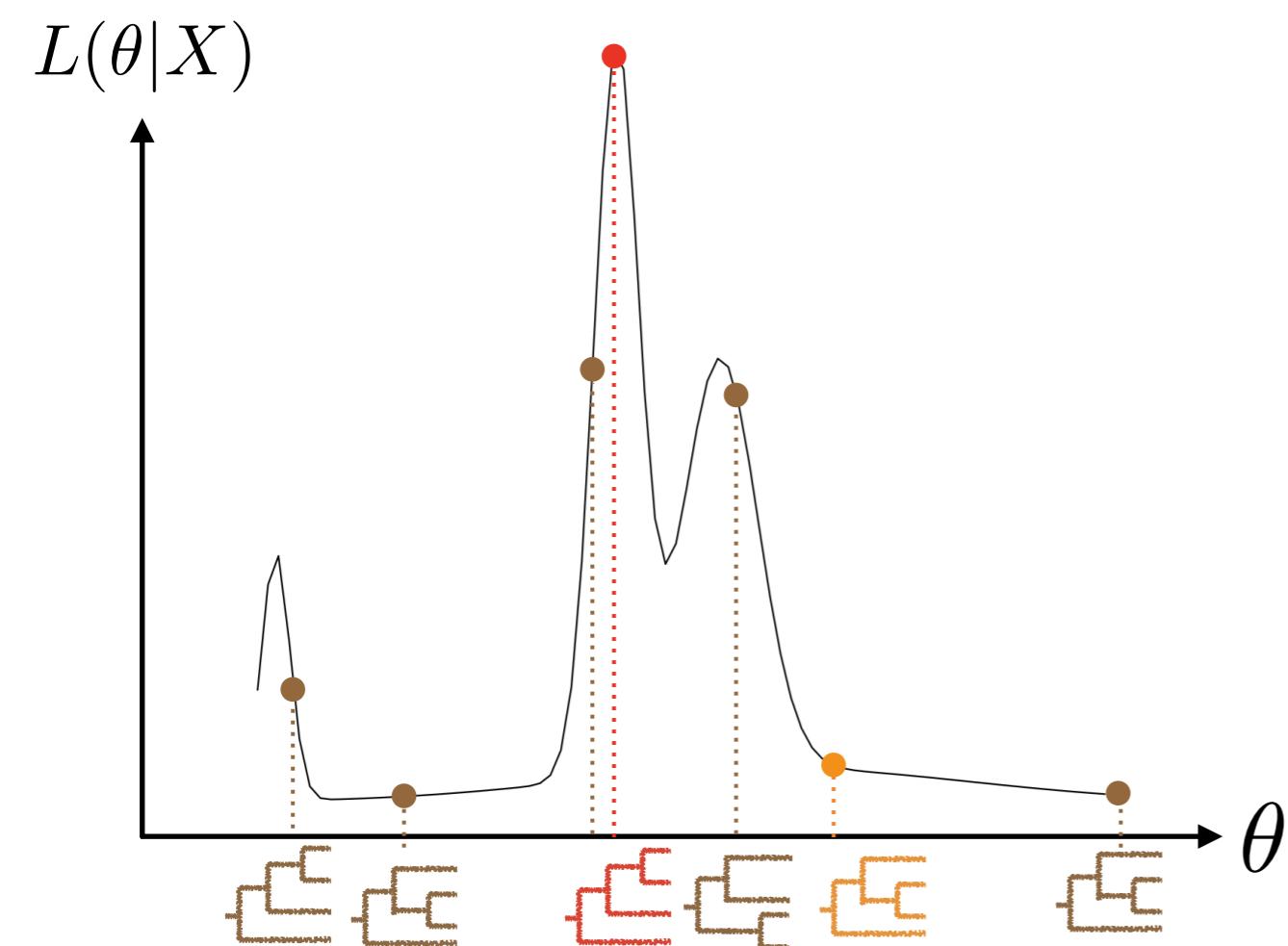
crsl4

(Solís-Lemus, Ané, 2016, PLoS Genetics)

Challenges

- Network space
- Identifiability
- Network comparison

Identifiability

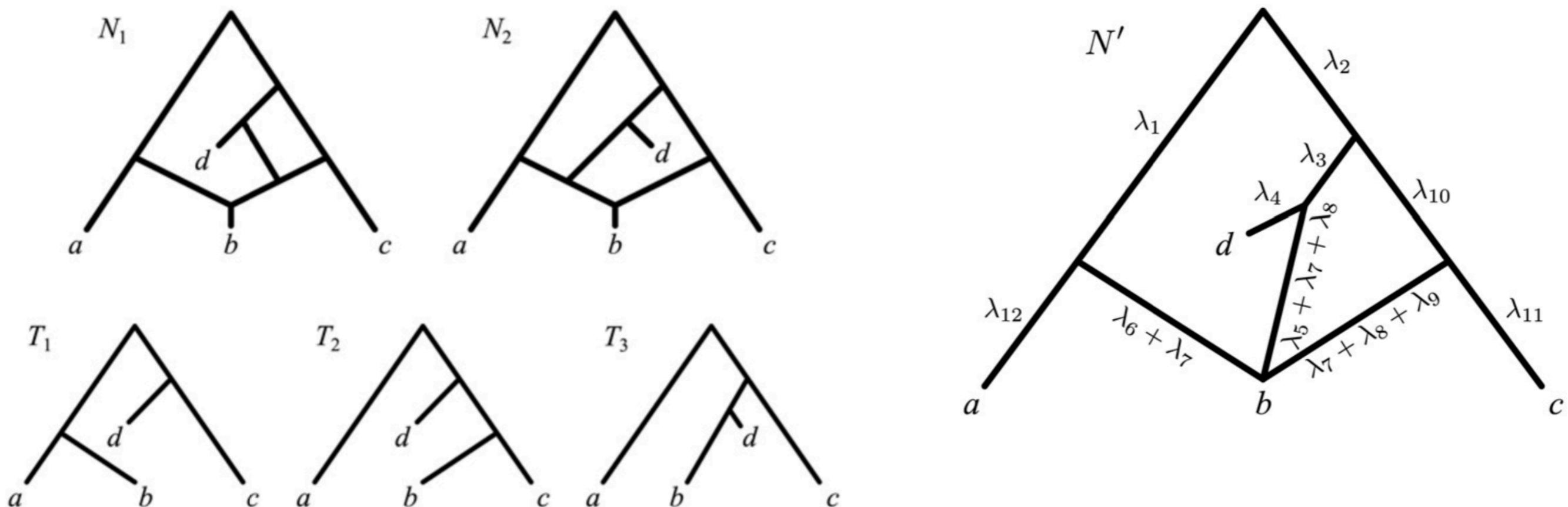


RESEARCH ARTICLE

Reconstructible Phylogenetic Networks: Do Not Distinguish the Indistinguishable

Fabio Pardi^{1,3*}, Celine Scornavacca^{2,3}

1 Laboratoire d’Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM, UMR 5506) CNRS, Université de Montpellier, France, **2** Institut des Sciences de l’Evolution de Montpellier (ISE-M, UMR 5554) CNRS, IRD, Université de Montpellier, France, **3** Institut de Biologie Computationnelle, Montpellier, France

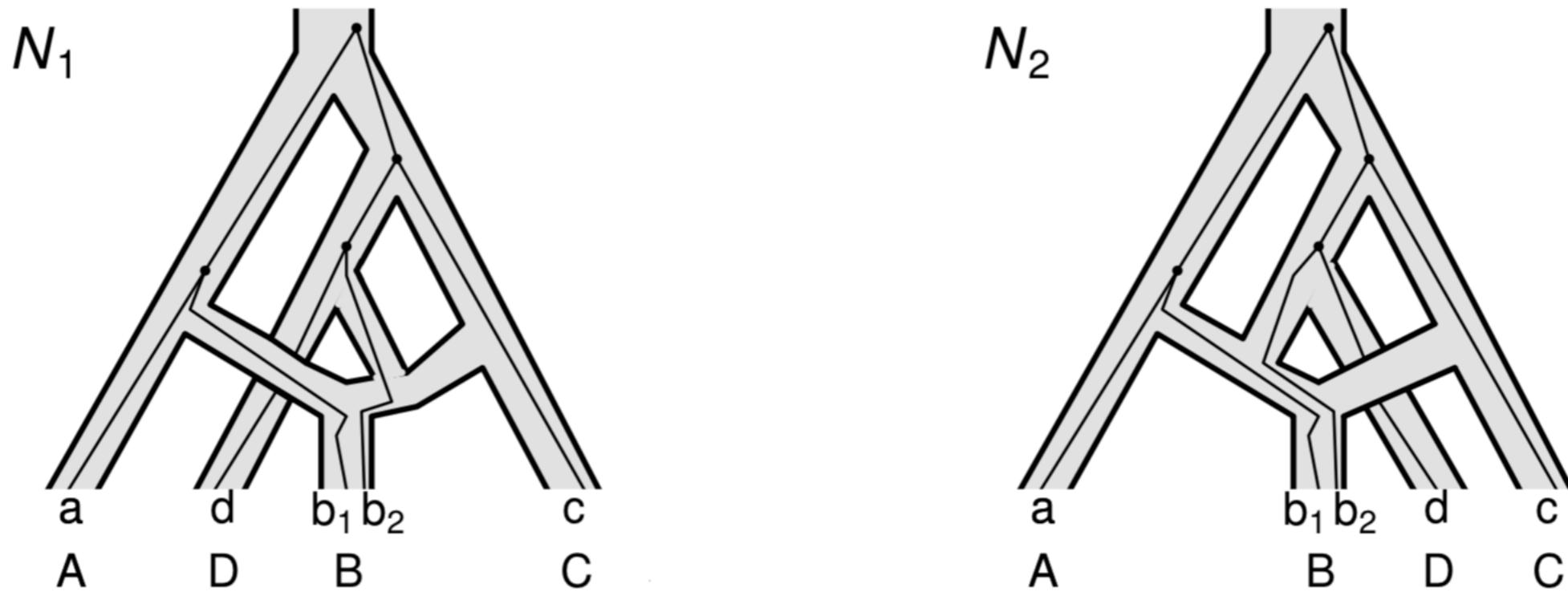


Undistinguishable with the
“displayed trees” criterion

Solution: Canonical
network (“unzipped”)

Displayed Trees Do Not Determine Distinguishability Under the Network Multispecies Coalescent

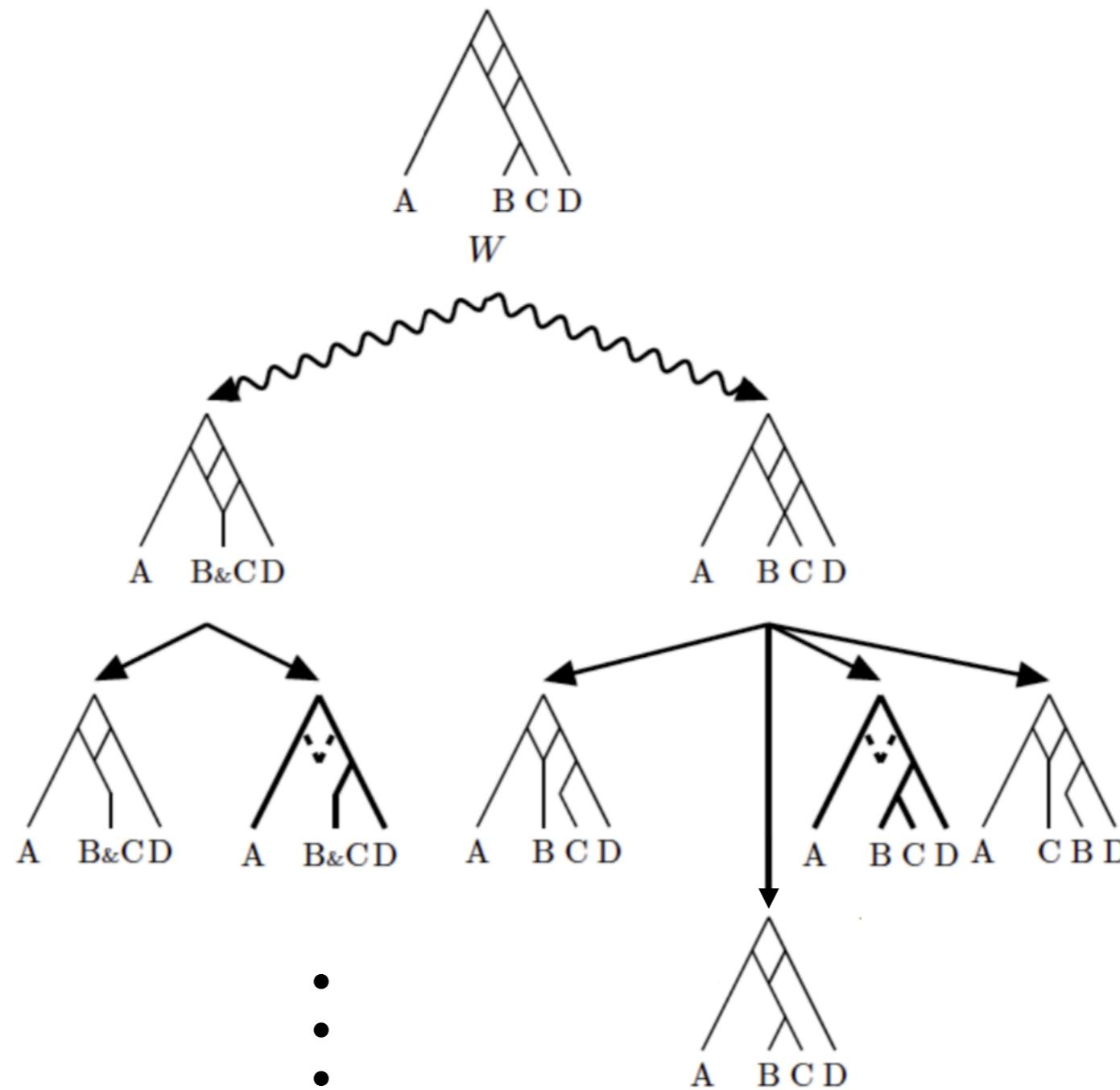
Sha Zhu¹, James H. Degnan²



Distinguishable under the MSC

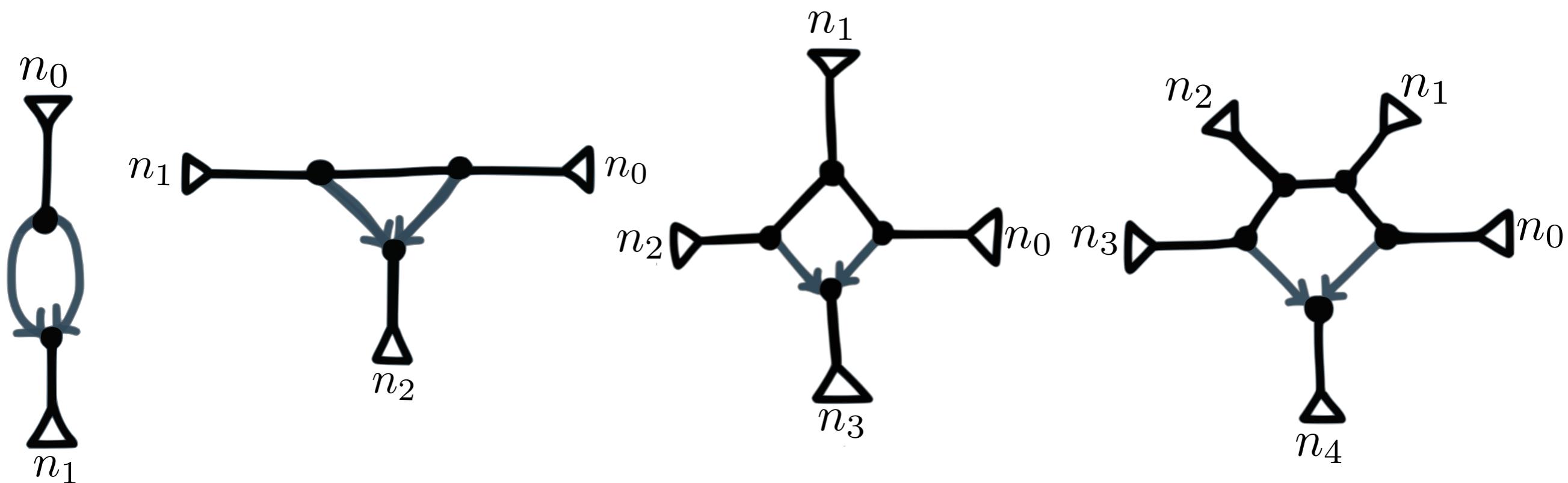
Displayed Trees Do Not Determine Distinguishability Under the Network Multispecies Coalescent

Sha Zhu¹, James H. Degnan²



Decomposing network in **parental** trees

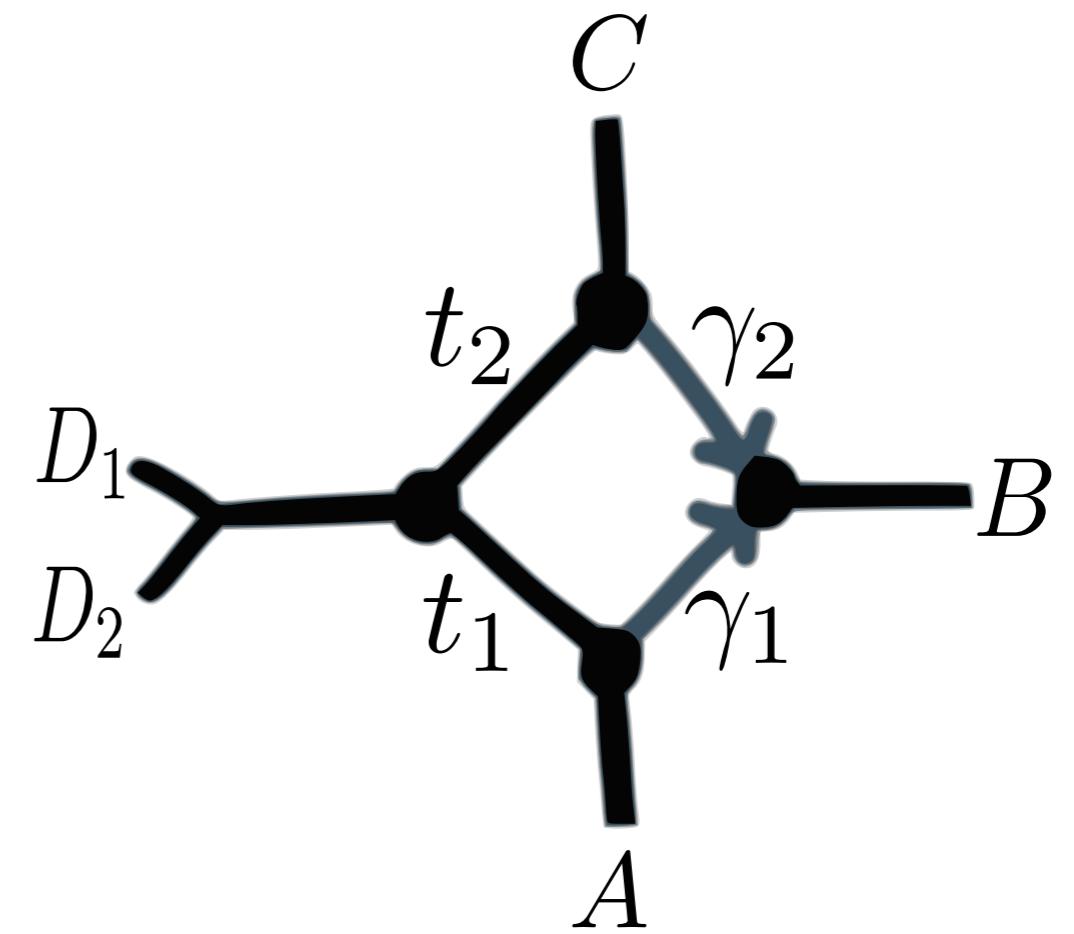
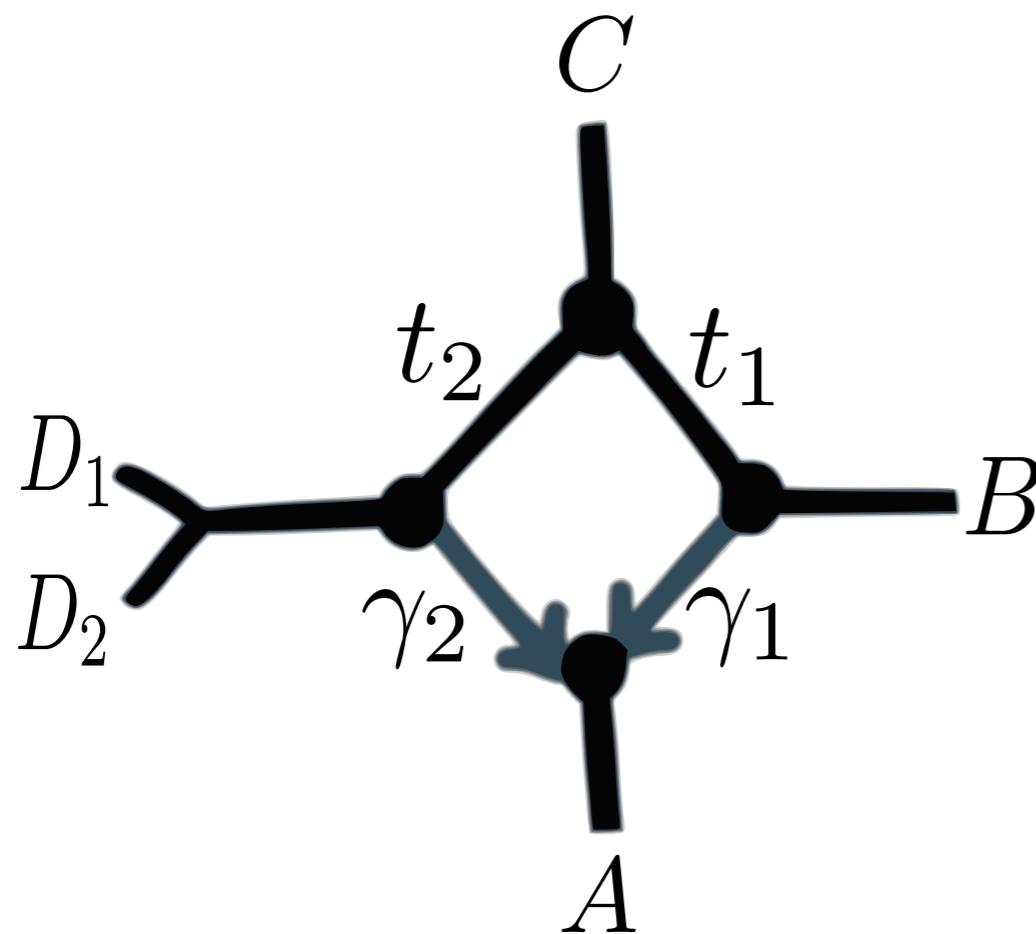
RESEARCH ARTICLE

Inferring Phylogenetic Networks with
Maximum Pseudolikelihood under
Incomplete Lineage SortingClaudia Solís-Lemus^{1*}, Cécile Ané^{1,2}Can we detect the
presence of
hybridization in level-1
networks?**No****Yes**
 $(n_i, n_j \geq 2)$ **Yes**
 $(n_i \geq 2)$ **Yes**

Generic Identifiability

 $t_i \in (0, \infty), \gamma \in (0, 1)$

In practice: flat pseudolikelihood

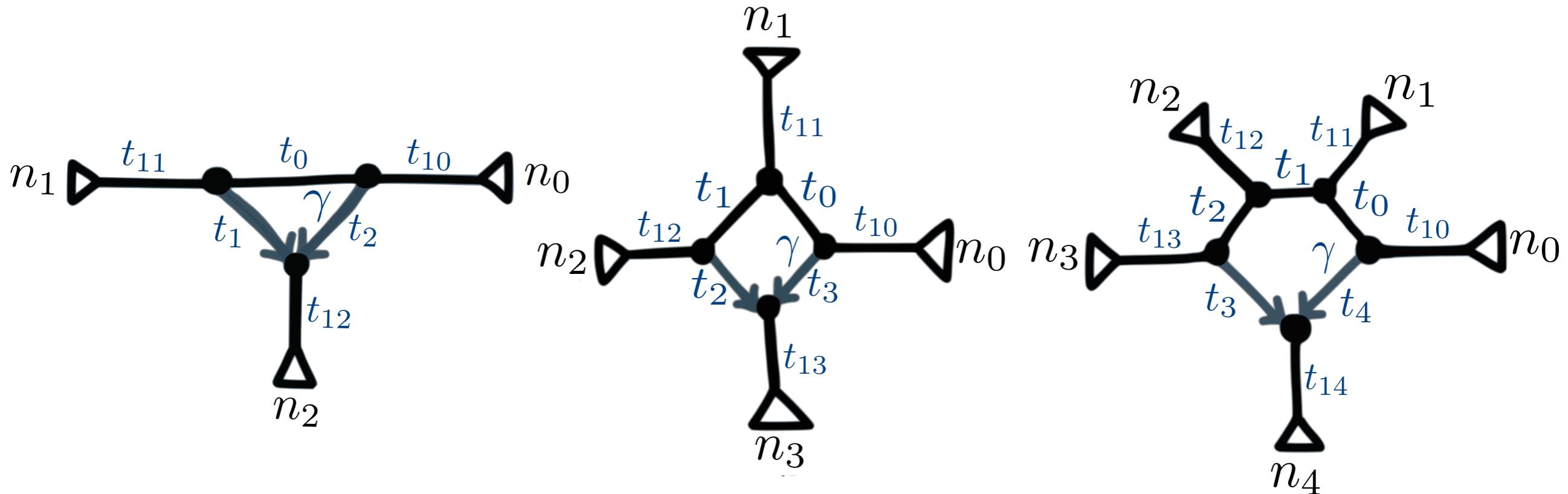


Can we estimate numerical parameters?

RESEARCH ARTICLE

Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting

Claudia Solís-Lemus^{1*}, Cécile Ané^{1,2}



No

Good triangle
($t_{12} = 0$)

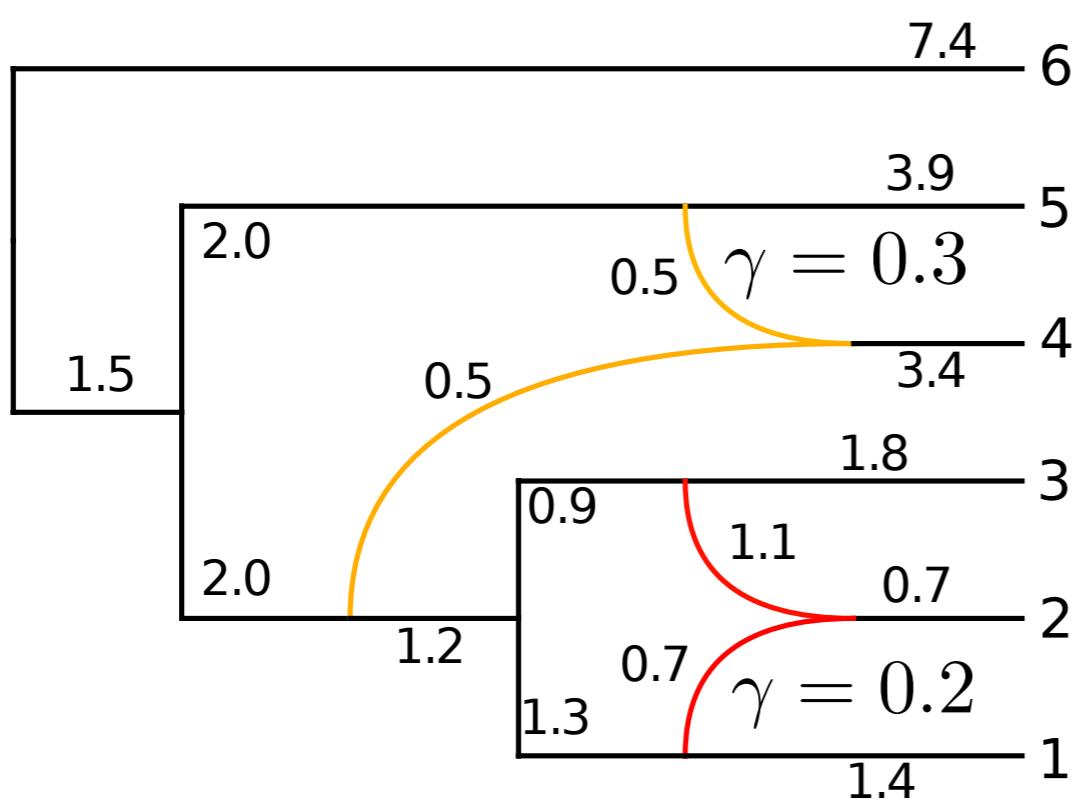
Yes

Good diamond
($n_0, n_2 \geq 2$)

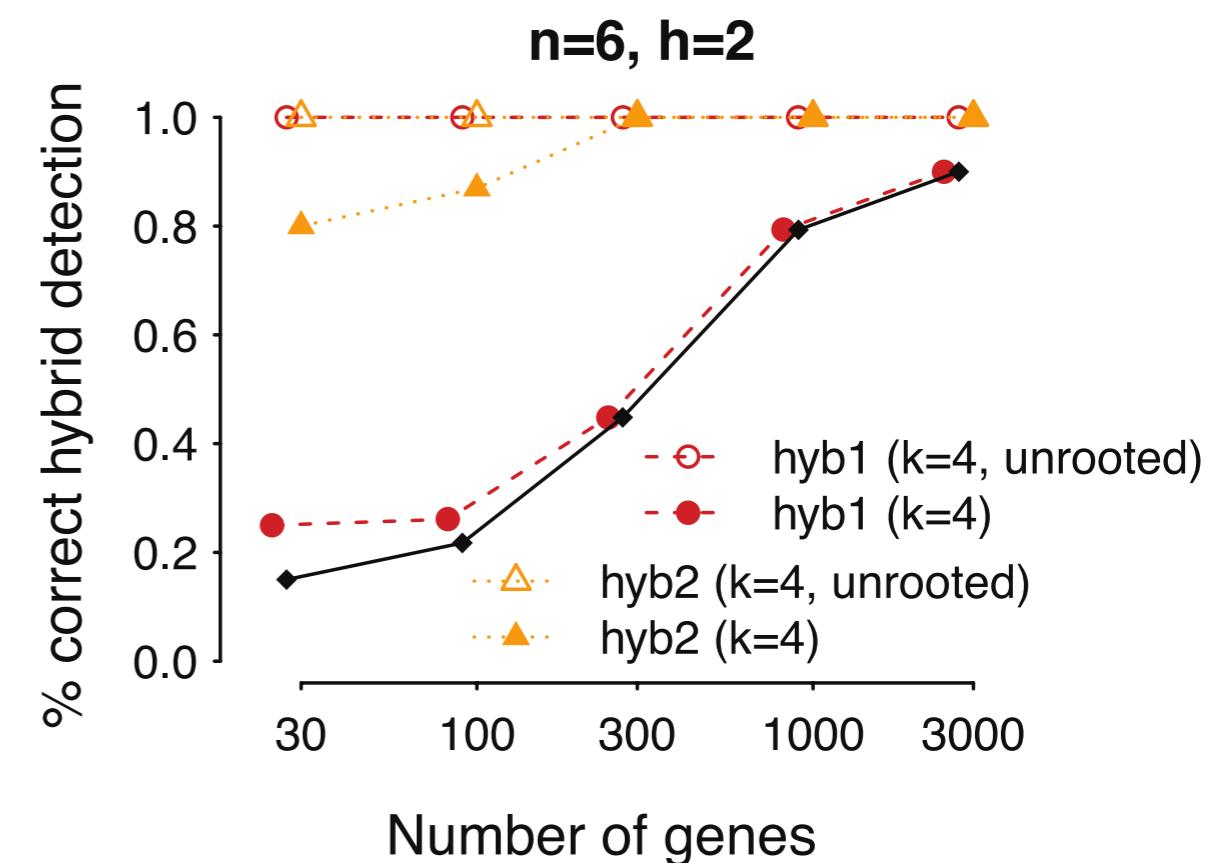
Generic Identifiability $t_i \in (0, \infty), \gamma \in (0, 1)$

Identifiability matters: SNaQ performance

Good diamond



Bad diamond



Challenges

- Network space

- Identifiability

Displayed vs Parental trees
Level-1 semi-directed networks
Hybridizations: case by case
Missing: likelihood, level-k semi-directed

- Network comparison

Challenges

- Network space

K. Huber, V. Moulton, C. Scornavacca,...
Missing: path through tree space, semi-directed

- Identifiability

Displayed vs Parental trees
Level-1 semi-directed networks
Hybridizations: case by case
Missing: likelihood, level-k semi-directed

- Network comparison

Challenges

- Network space

K. Huber, V. Moulton, C. Scornavacca,...
Missing: path through tree space, semi-directed

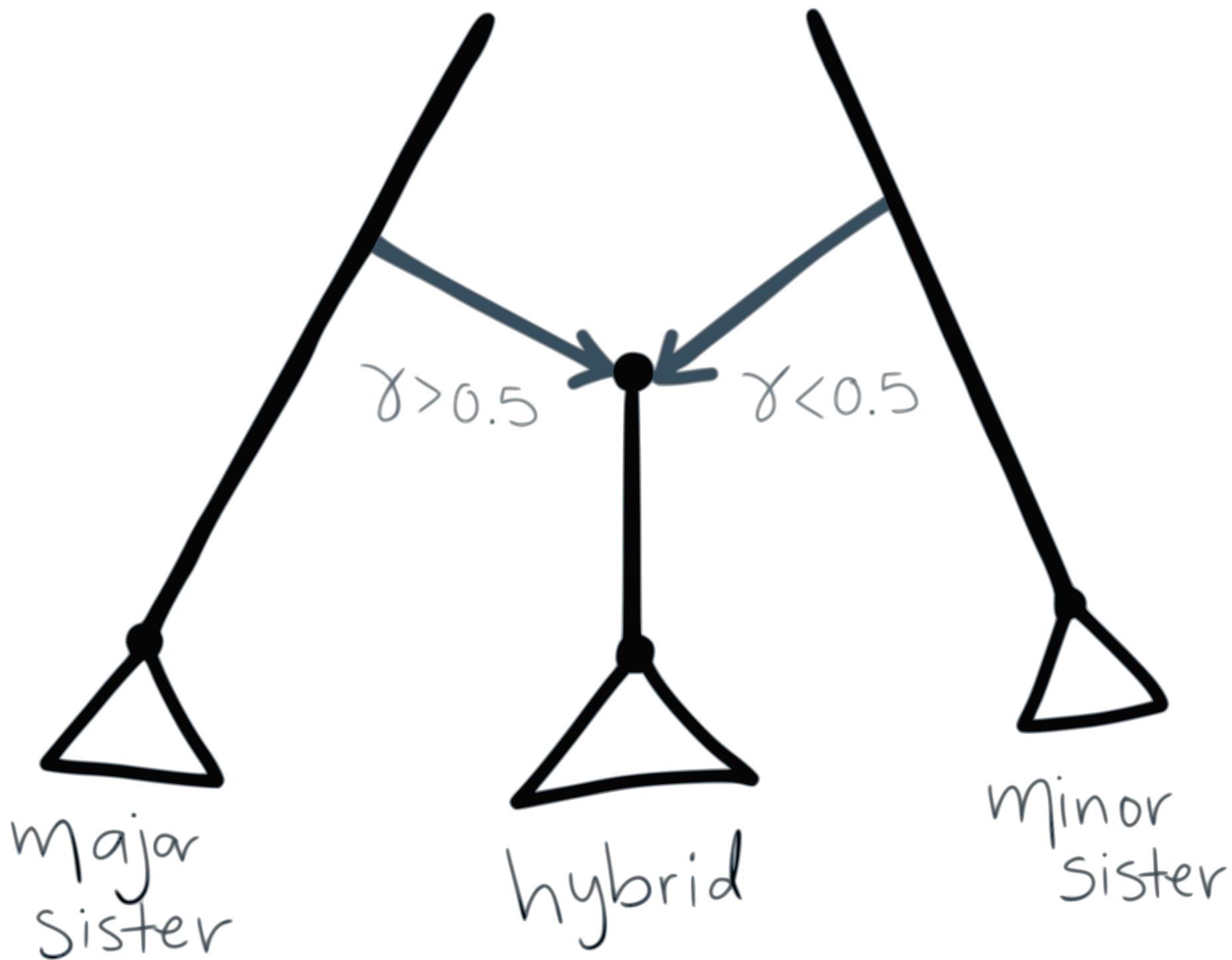
- Identifiability

Displayed vs Parental trees
Level-1 semi-directed networks
Hybridizations: case by case
Missing: likelihood, level-k semi-directed

- Network comparison

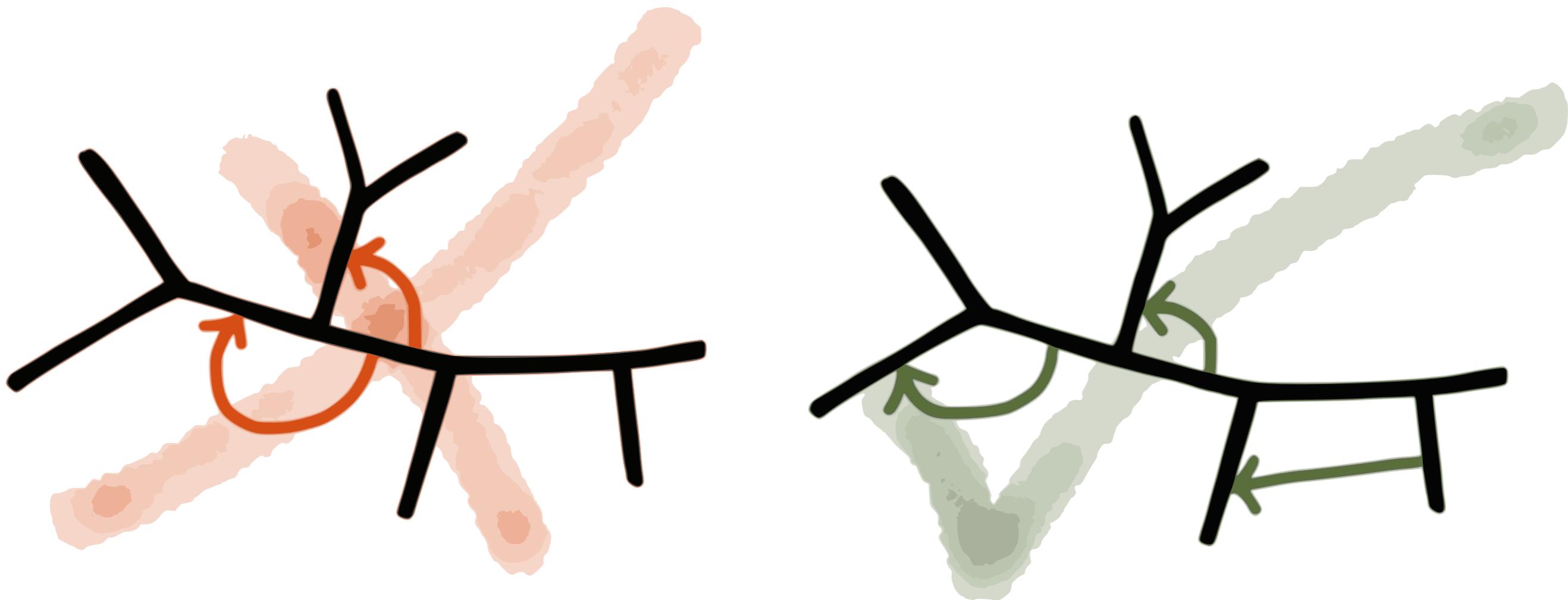

Missing: distance function
Hardwired-cluster distance only for rooted networks
Summary of networks: clades!

Network summary



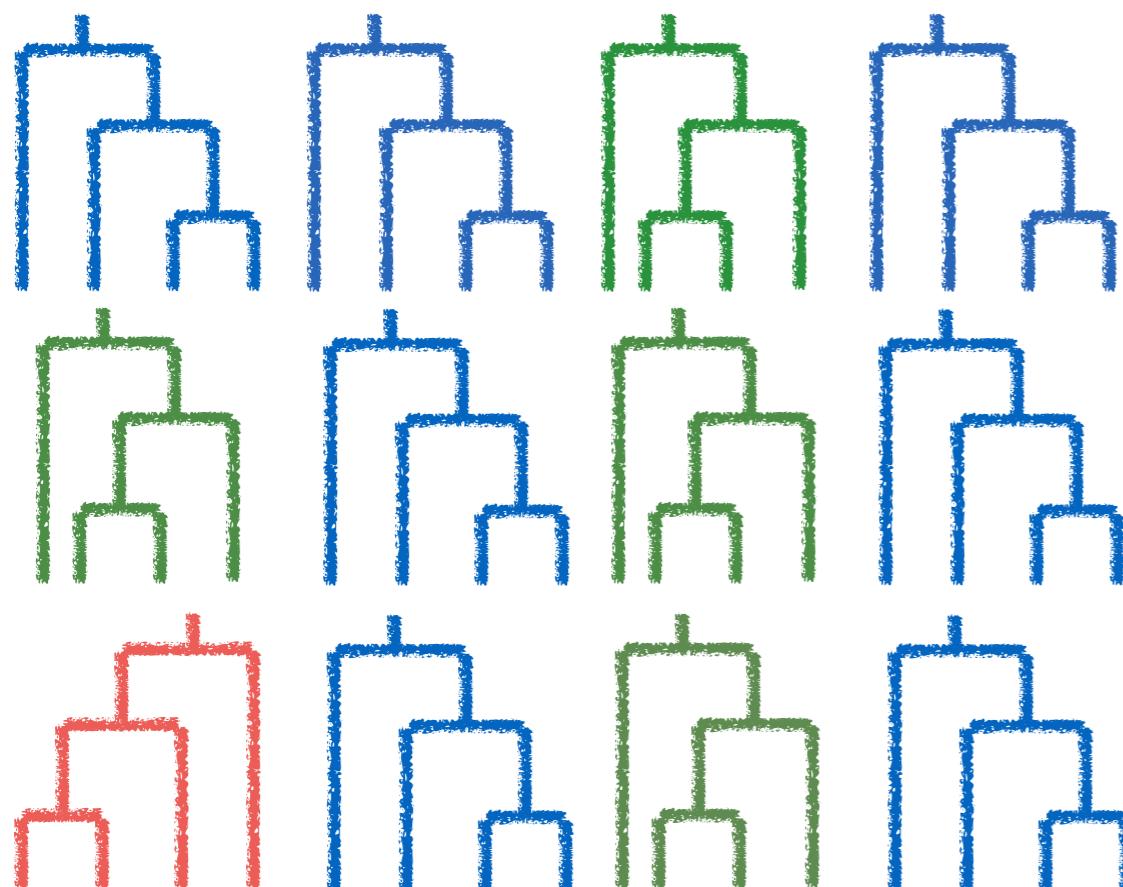
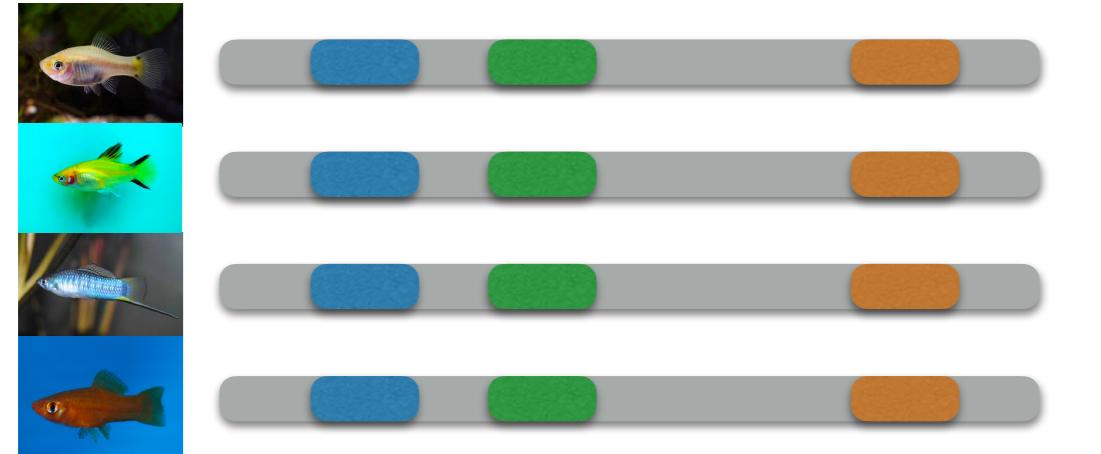
(S.-L. et al, 2017, MBE)

snaQ limitation: Level-1 networks

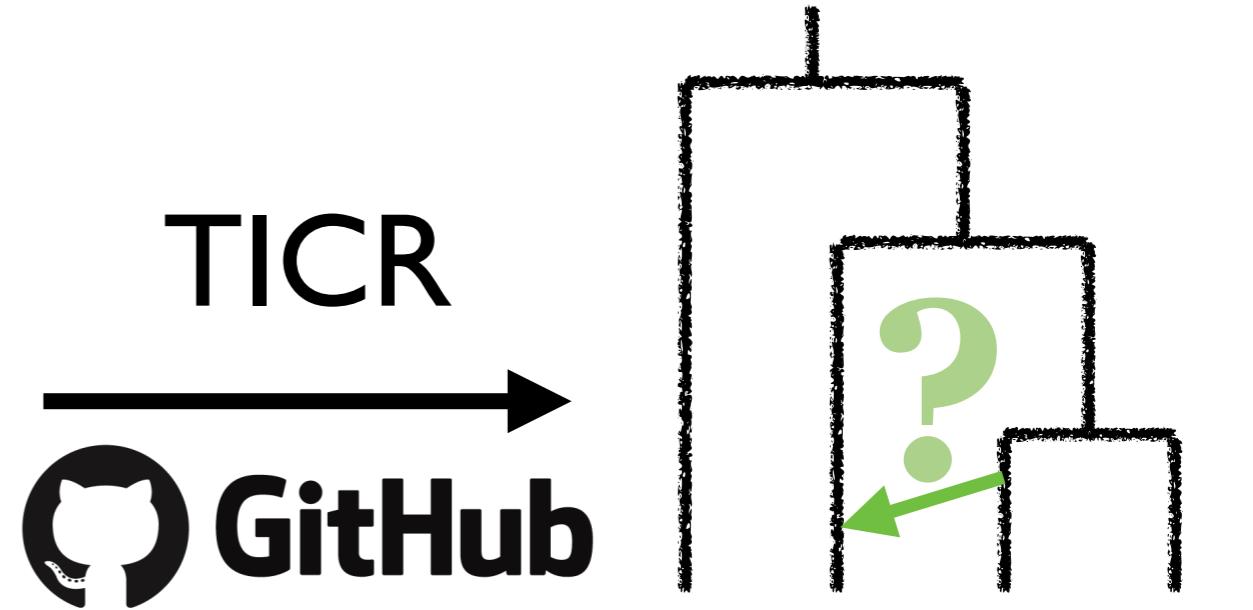


When?

Phylogenetic network



Goodness-of-fit test
Hypothesis test:
Is a tree a good fit?



<https://github.com/nstenz/TICR>
(Stenz et al, 2015, Syst Bio)

PhyloNetworks: analysis for phylogenetic networks

build passing docs stable docs dev codecov 81% coverage 67%

Overview

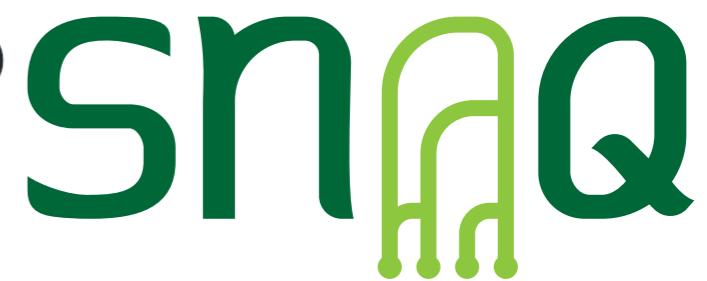


PhyloNetworks is a [Julia](#) package with utilities to:

- read / write phylogenetic trees and networks, in (extended) Newick format. Networks are considered explicit: nodes represent ancestral species. They can be rooted or unrooted.
- manipulate networks: re-root, prune taxa, remove hybrid edges, extract the major tree from a network, extract displayed networks / trees
- compare networks / trees with dissimilarity measures (Robinson-Foulds distance on trees)
- summarize samples of bootstrap networks (or trees) with edge and node support
- estimate species networks from multilocus data (see below)
- phylogenetic comparative methods for continuous trait evolution on species networks / trees



- Step-by-step tutorial
- Online documentation
- Google user group



(S.-L. et al, 2017, MBE)



<https://solislemuslab.github.io/>



@solislemuslab



crsl4

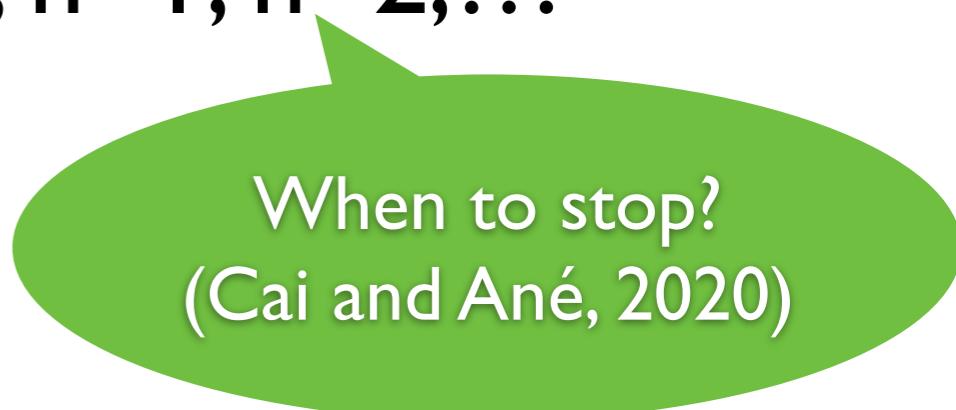
Practical advice

- Do multiple runs
- Do bootstrap
- Check the .networks output file (especially if hybridization conflicts with outgroup)
- What is the quality of my input data (gene trees/CFs)?
- Run SNaQ sequentially: $h=0, h=1, h=2, \dots$



Practical advice

- Do multiple runs
- Do bootstrap
- Check the .networks output file (especially if hybridization conflicts with outgroup)
- What is the quality of my input data (gene trees/CFs)?
- Run SNaQ sequentially: $h=0, h=1, h=2, \dots$



When to stop?
(Cai and Ané, 2020)



In-class dynamic

- **Time:** 20 minutes
- **Instructions:** We will go over the PhyloNetworks pipeline which will cover ASTRAL, BUCKy and SNaQ. Create our own reproducible script.